IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 18, 2018, accepted October 8, 2018, date of publication October 22, 2018,
date of current version December 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877401

Mask Optimization for Image Inpainting

MARIKO ISOGAWA[1,3], DAN MIKAMI[1,2], (Member, IEEE), DAISUKE IWAI[3], (Member, IEEE), HIDEAKI KIMATA[1], AND KOSUKE SATO[3], (Member, IEEE)

[1]NTT Media Intelligence Laboratories, Yokosuka 239-0847, Japan
[2]NTT Communication Science Laboratories, Atsugi 243-0198, Japan
[3]Graduate School of Engineering Science, Osaka University, Toyonaka 560-8531, Japan

Corresponding author: Mariko Isogawa (isogawa@sens.sys.es.osaka-u.ac.jp)

ABSTRACT This paper proposes a novel approach to image inpainting that optimizes the shape of masked regions given by users. In image inpainting, which removes and restores unwanted regions in images, users draw masks to specify the regions. However, it is widely known that the users typically need to adjust the masked region by trial and error until they obtain the desired natural inpainting result, because inpainting quality is significantly affected by even a slight change in the mask. This manual masking takes a great deal of users' working time and requires considerable input. To reduce the human labor required, we propose a method for masked region optimization so that good inpainting results can be automatically obtained. To this end, our approach estimates ''naturalness of inpainting'' for all super pixels in inpainted images and reforms an original mask on a super-pixel basis, so that the naturalness of the inpainting result is improved. The efficacy of this approach does not depend on inpainting algorithms, thus it can be applied for every inpainting method as a plug-in. To demonstrate the effectiveness of our approach, we test our algorithm with varied images and show that it outperforms the existing inpainting methods without masked region reformation.

INDEX TERMS Inpainting, super pixel, learning-to-rank, segmentation.

I. 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 has been researched to automatically remove such unwanted regions and fill them in as ''perceptually-naturally'' as possible. Patch-based inpainting [1]–[5] is acknowledged as a promising approach that provides perceptually acceptable inpainting quality, and thus has been applied in many commercially available software packages. Recently, researchers have tried to apply convolutional neural networks (CNNs) to this domain, and in doing so showed great improvement could be obtained in the inpainting performance [6]–[10]. However, all the inpainting algorithms reported so far share the same limitation. They assume that a user manually draws a mask to specify an unwanted region as an offline process; they mask either by drawing the boundary [11]–[13] or marking points on the target region [14], [15]. The inpainting quality thus significantly depends on the masked region.

To explain this bottleneck and achieve better inpainting, we go back to the workflow of inpainting with the basic patch-based approach. It consists of two parts: (1) users mask unwanted regions in the input image and (2) users execute the inpainting process for the masked image. Existing inpainting methods focus only on the latter part.

To review how the masked region affects inpainting quality, we consider an inpainting task to remove the image of a woman standing in front of the image of another woman with two different masked regions as shown in Figure 1. Hereinafter, we focus on the patch-based approach, while the discussion here does not lose the generality. In Figure 1, both (1-a) and (1-b) and (2-a) and (2-b) represent a masked region and its inpainted results. Interestingly, although it is typically expected that better inpainting results should be obtained for (1-a) because the masked region is smaller, they are actually less natural.

The reason differences due to the masked region occur is shown by blue patches in Figure 1. The patch-based approach divides the original image into small patches for replacing a patch containing masked region with a similar patch only containing source region. Note that the source region consists of pixels that do not belong to the masked region. The similarity is computed merely by using the source region in the patch. In the image shown in (1-a) there are no similar patches to the blue patch since the latter includes a unique texture, i.e., that of a part of the neighboring person's body.

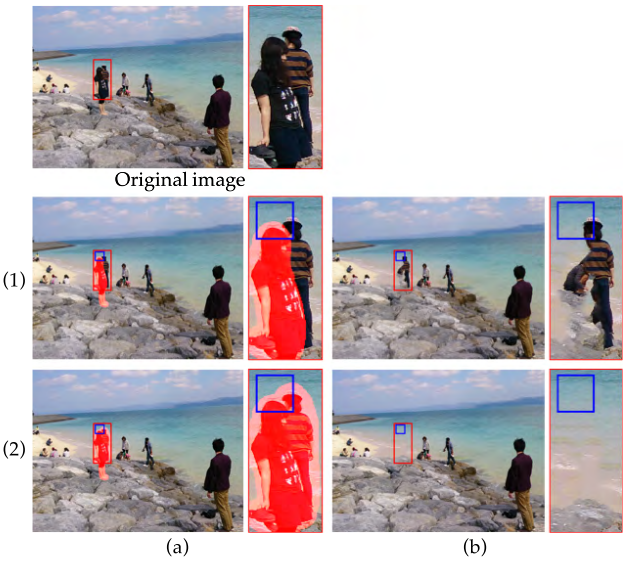


FIGURE 1. An example that shows how masked region reformation affects total inpainting quality. Although the inpainted result is unnatural with the original masked region (1), the reformed masked region (2) achieves a better inpainted result. (a)masked image. (b) Inpainted image.

Patch retrieval failure deteriorates inpainting quality as shown in Figure 1(1-b). In contrast, with the blue patch in (2-a), which includes only sea textures outside of the mask textures, many similar patches exist. Thus, the inpainting quality of (2-b) becomes better though the source region is smaller. As shown in this example, the masked region should be optimized to achieve perceptually natural inpainting results. Although one might think the best mask region could be acquired by perfectly segmenting the object region, a desired inpainting result is not always generated in the manner discussed above.

This paper proposes a masked region optimization method. The method reforms an initial masked region drawn by a user towards perceptually natural inpainted results. In particular, the method iteratively localizes unnaturally inpainted regions and reforms the masked region so that the localized regions do not form the contour of masked region. There are two technical issues to achieve this: (1) localizing unnaturally inpainted region, and (2) reforming the masked region.

For the former issue, we consider applying an image quality assessment (IQA) technique for image inpainting [16]. The original method applies a learning-to-rank approach to judge which inpainting result is more natural given two inpainted results. However, it does not localize unnatural regions (in this paper we define ''unnatural'' as ''unnatural to human perception due to inpainting failure'') in all images and does not indicate where the masked region should be reformed. Thus it cannot be directly applied to our method. For the latter issue, we should consider an efficient strategy for reforming masked regions regarding computational cost. Since unnatural region localization should be performed for all masked regions, a huge number of iterations (masked region reformation, inpainting, and assessment) would be required if we reformed the masked region on a per-pixel basis.

Therefore, to address both the former and latter issues, we applied the super pixel concept [17]. This is an entity that groups similar pixels given specified criteria. It can be used as computation units of localized unnaturalness and reformation to reduce the computational cost. In the work we report in this paper, we extended IQA methods for inpainting so that they locally assess the naturalness of super pixels rather than entire images. Then we dilated or eroded masked regions so that the super pixels with unnaturalness do not form the contour of the masked region.

Contributions: The main contribution of this paper is proposing masked region optimization, a new solution to improve image inpainting. We also propose super-pixel-wise unnatural region localization and masked region reformation algorithms to verify the validity of the main contribution.

The rest of this paper is organized as follows. In Section **II** we briefly review related work. Section **III** describes the super-pixel-based mask optimization approach we propose. In Section **IV**, we verify the method's efficacy with varied experiments. Section **V** reviews and discusses experimental results, and in Section **VI**, we conclude the paper with a summary of key points and a mention of future work to be done.

II. RELATED WORK

This section first introduces existing work for image inpainting in subsection II-A. Then, we review existing methods to solve the two remaining issues we described in the previous section. Subsection II-B shows IQA methods for naturalness estimation, and then subsection II-C introduces super pixels as the reformation unit.

A. IMAGE INPAINTING METHODS

Many effective approaches for image inpainting have been proposed [18]. Exemplar-based inpainting methods that fill unwanted regions by using other regions in images or databases are acknowledged as a promising approach. Criminisi *et al.* [1] first proposed exemplar-based inpainting based on patch retrieval. This algorithm replaces a target patch with patches similar to it from source regions in images. However, this method fails if there are no appropriate patches in the source region. To overcome the limitation, approaches that increase the patch availability have been proposed. There are mainly two approaches: transforming patches or retrieving patches with relaxed constraints. Patch transforming approaches use patches unsuitable for filling holes in their original condition by transforming the patch geometry. Darabi *et al.* [3] introduce scaling and rotation of patches while Huang *et al.* [5] allow projective transformation. As an approach with relaxed constraints, it was found that retrieving patches in different feature space makes restoration more effective than in original spaces such as motion field [19] and lower dimensional space [20].

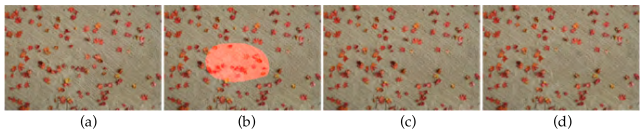


FIGURE 2. An example that shows the difficulty in evaluating inpainted images objectively. (a) and (b) are the original image and the masked region. (c) and (d) are inpainted images for (b) with different algorithms. Although both results are different from the original, they are perceptually natural. (a) Original image. (b) Original image with masked region. (c) Inpainted image 1. (d) Inpainted image 2.

CNN-based inpainting methods have also been proposed [6]–[8]. Yang *et al.* [10] extended CNN-based inpainting to larger masked regions. They proposed a context encoder to learn features by inpainting based on GAN. Iizuka *et al.* [9] proposed locally and globally consistent inpainting based on GAN. To train the network, they use global and local context discriminators to distinguish real images from completed ones. Although many effective methods have been proposed as described above, no inpainting algorithms have shown to be successful if the masked region is not appropriate for the inpainting task. In addition, even subtle changes in masked regions generate huge differences in inpainted results as we showed in Figure 1. In our work, we overcome this bottleneck by optimizing masked regions.

B. IQA METHODS FOR INPAINTED IMAGES

Assessing naturalness of inpainted images is acknowledged as a task that can only be done by subjective judgment. One primary reason is explained by using Figure 2. In the figure, (a) and (b) respectively show an original image and one with a masked region, while (c) and (d) are inpainted images for (b) with different algorithms. Although both of these results are different from the original image, these results are perceptually natural. In the inpainting context, these results are "correct."

Existing IQA methods have tried to find a way to represent subjective quality of naturalness of inpainted results by means of objectively measurable indicators. Venkatesh and Cheung used observed gaze density inside and outside the masked region in inpainted images [21]. Instead of observed gaze, many IQA methods use a computational visual saliency map, which simulates human gaze density [22]–[26]. However, actual human gazes vary for individual viewers and viewing contexts and their correspondence with saliency maps is quite limited. Some recent saliency models are robust to general image degradation factors such as blurring, downsampling, or compression noise [27], [28]. However, they are not dedicated for finding unnaturalness in inpainted images.

Thus, to estimate such unnaturalness, machine learning based IQA methods have been developed [16], [25], [26]. Oncu *et al.* [25] and Trung *et al.* [26] proposed support vector regression (SVR)-based approaches. Boykov and Jolly *et al.* [16] achieved more accurate subjective unnaturalness estimation for inpainted images by dividing the problem into a set of pairwise preference order estimation tasks and using the learning-to-rank approach, whose concept has been widely applied (not limited to image quality) to various tasks requiring subjective judgments [29]–[32].

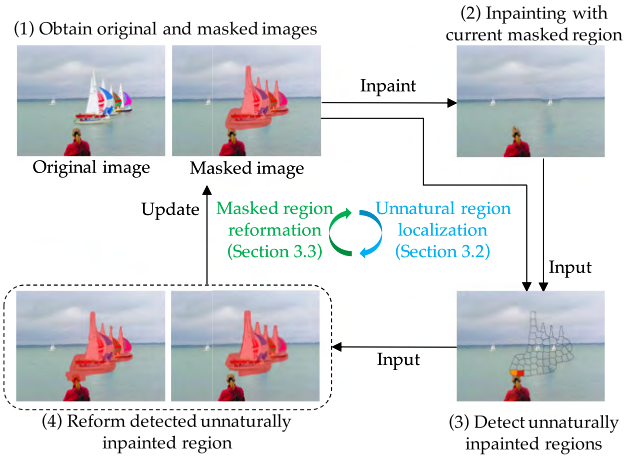


FIGURE 3. Proposed mask optimization framework overview. The method consists of four steps: (1) obtain an original and a masked image, (2) inpaint with current mask, (3) localize unnaturally inpainted region, and (4) reform current masked region. The procedure is repeated until Step 3 does not detect any unnaturally inpainted regions.

The method focuses on estimating preference orders rather than absolute scores. Here, the preference orders represent which inpainted images are more preferred (i.e., natural) by human perception. Preference orders allow us to select the best one from multiple inpainting results. The important advantage of a learning-to-rank-based approach is that it can learn only on the basis of rank order. In our work we used this learning-to-rank-based IQA method [16] as an optimization indicator.

C. SUPER PIXEL

The "super pixel" concept, originally developed by Vezhnevets and Konouchine *et al.* [17], is a perceptually meaningful entity that groups similar pixels into smaller regions. Such super pixels have many desired properties. By grouping the pixels, super pixels reduce computational complexity [33]. They also reduce processing complexity; they carry more information than pixels and thus are perceptually meaningful objects, having the scale between the pixel level and the object level [34], [35].

Currently many applications have been proposed on the basis of such properties of super pixels [34]–[36]. Lucchi *et al.* [37] use super pixels for image segmentation to reduce computational cost and enforce local consistency. Super pixels have also been effectively used with tracking tasks [38], [39]. In such cases super pixels are used as the perspective representation of mid-level features. Wang *et al.* [40] also proposed a super-pixel-based graphical model for remote sensing. They introduced super pixels as new basic units in conditional random field modeling.

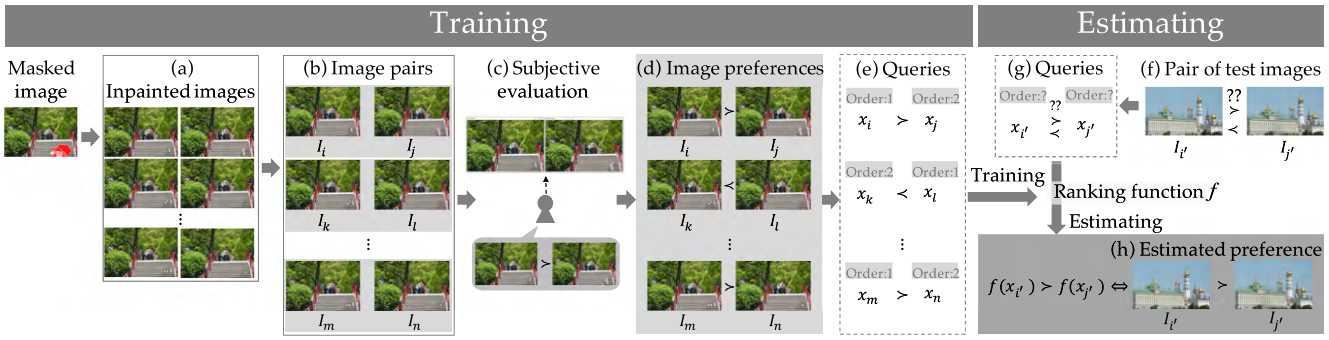


FIGURE 4. Overview of IQA method for inpainted image [16] that our unnatural region detection uses.

To achieve computational and processing efficacy, we apply super pixels as a unit for unnatural region estimation and masked region reformation. Section III explains this in more detail.

III. PROPOSED METHOD

We propose a mask optimization method for image inpainting. The key idea is that a masked region is reformed so that the regions that are unnaturally inpainted do not form the contours of the masked region. The proposed method consists of the following four steps (see Figure 3) and the procedure is repeated until Step 3 does not detect any unnatural regions.

- **Step 1** Obtaining the original image and the current masked image (manually designated or updated after Step 4)
- **Step 2** Inpainting the current masked image
- **Step 3** Localizing the unnaturally inpainted regions
- **Step 4** Reforming the mask depending on localized unnatural regions

So far, our algorithm only supports reformation in one direction, i.e., dilation or erosion. Neither larger nor smaller masked regions beyond those that are necessary decrease inpainting quality. Larger masked regions may overlap neighbor objects and reduce source regions used for filling holes. Smaller masked regions may reveal target objects that are desired to be removed.

For proposed masked region optimization, we need to address two issues. One is a way to localize unnaturally inpainted regions in Step 3. The other is a way to reform the masked region in Step 4. The following subsections first introduce a previous IQA method for inpainted images [16] in III-A, since it is the key method for our proposed method. Then, III-B and III-C respectively describe localization of unnatural regions and masked region reformation.

A. LEARNING-TO-RANK BASED IMAGE QUALITY ASSESSMENT

Before we describe our proposed method's details, this subsection introduces the learning-to-rank-based quality assessment for inpainted images [16], which is used in developing our unnatural region localization. This method premises a ranking function $f(x)$ that projects inpainted images to a one-dimensional axis in accordance with unnatural inpainting.

The overview of the method's framework is shown in Figure 4. As training data, paired inpainted images are obtained (see Fig. 4(b)) with several inpainted images I^i with varied parameters as shown in Fig. 4(a). Then, subjective preference orders are manually annotated (see Fig. 4(c)) to generate inpainted pairs with preferences (see Fig. 4(d)). These samples are extracted into feature vectors x^i as shown in Fig. 4(e) to train ranking function $f(x)$. Given two inpainted images as shown in Fig. 4(f) with their extracted features (see Fig. 4(g)), preference orders for these two inpainted images are obtained as output values via $f(x)$. The training and estimation processes are explained below.

1) TRAIN THE RANKING FUNCTION

Hereafter, we use "$I^i \succ I^j$" to express that "I^i is preferred to I^j". We define the function $h(x^i, x^j)$ that denotes annotated preferences by subjects as follows:

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

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

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

The method models $f(x)$ with the linear function $f(x) = \omega^\top x$. Accordingly, Eq. 2 can be rewritten as

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

Then, the weight vector ω satisfying Eq. 3 for most training data pairs is found. This is the same problem as that of binary classification. The method uses a pairwise learning-to-rank algorithm called RankingSVM [41] to solve it.

2) ESTIMATE PREFERENCE ORDERS

Given pair-wise inpainted images I^a and I^b with their image feature vectors x^a and x^b, output of ranking functions $f(x^a)$

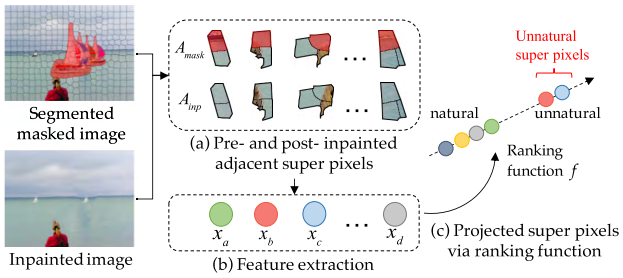


FIGURE 5. Super pixel projection to eigenspace that represents inpainting unnaturalness. With pre- and post- inpainted super pixels (a), feature vector of these super pixels are computed as shown in (b). These feature vectors are projected into an eigenspace via ranking function (c). Outlier samples in the space are detected as unnatural super pixels.

and $f(x^b)$ are calculated for all images. The preference orders between I^a and I^b are obtained as $I^a \succ I^b$ when $f(x^a) > f(x^b)$, and $I^b \succ I^a$ when $f(x^b) > f(x^a)$.

In contrast to this previous IQA method for a single **whole image** evaluation, we need a method to **localize** unnaturally inpainted regions toward a better mask. The next subsection describes how our proposed method accomplishes this.

B. ESTIMATING UNNATURAL REGION WITH LEARNING-TO-RANK

Now we are ready to explain how we localize unnatural regions in inpainted images. We have added two improvements to the previous IQA method. First, the proposed method evaluates the relative unnaturalness for the super pixels instead of the whole images by considering a pair of adjacent super pixels as a single image. Second, it localizes unnatural super pixels by finding outlier ones in projected eigenspace that represent inpainting unnaturalness.

1) APPLYING IQA METHOD TO ADJACENT SUPER PIXELS

We applied the previous IQA method [16] to adjacent super pixel A. We denote this adjacent super pixel as A^{mask} for a masked image and A^{inp} for an inpainted image. A^{mask} is composed of the contours of the masked region. It consists of two super pixels, one in the masked region and the other at the outside of the masked region. We denote the super pixels in the masked region as S^{mask}. Similarly, A^{inp} consists of one inpainted super pixel and the other at the outside of the inpainted region.

By considering A^{mask} and A^{inp} as pre- and post- inpainting images as shown in Figure 5(a), feature vectors of inpainted adjacent super pixel A^{inp} can be extracted as shown in Figure 5(b). Then, these features can be projected via a ranking function into a one-dimensional axis representing inpainting unnaturalness (see Figure 5(c)). The feature vector calculation formula is provided in Appendix VI.

2) LOCALIZING UNNATURAL REGIONS

The ranking function behavior shows that nearby coordinates are mapped on the one-dimensional axis for similar samples. That is, the mapping via ranking function can be used to find outliers. Under the assumption that a majority of super pixel samples are naturally inpainted, outliers can be considered as unnaturally inpainted samples (see Figure 5(c)). Masked super pixels related with these outlier samples are detected as unnatural super pixels to be reformed.

Algorithm 1 Unnatural Super Pixels Localization

Input: Adjacent super pixels $\boldsymbol{A^{inp}}$ and corresponding masked super pixels $\boldsymbol{S^{mask}}$
Output: Unnatural super pixels $\boldsymbol{S^{ref}}$

$N = |\boldsymbol{A^{inp}}|$
for $n = 1$ to N **do**
  $x_n \Leftarrow CalculateFeatureVector(A_n^{inp})$
  $f(x_n) \Leftarrow OutputRankingValue(x_n)$
end for
$\boldsymbol{F(X)} = \{f(x_1), f(x_2), ..., f(x_N)\}$
if TH has not been calculated yet **then**
  $TH = (min(\boldsymbol{F(X)}) + max(\boldsymbol{F(X)}))/2$
end if
for $n = 1$ to N **do**
  if $f(x_n) < TH$ **then**
    Add S_n^{mask} to $\boldsymbol{S^{ref}}$
  end if
end for

Detailed algorithms are as follows (See Algorithm 1). Among all adjacent super pixels in inpainted image $\boldsymbol{A^{inp}} = \{A_1^{inp}, A_2^{inp}, ..., A_N^{inp}\}$ projected into an eigenspace via the ranking function, we find outlier adjacent super pixels. N is the amount of $\boldsymbol{A^{inp}}$. To find such outlier samples, a threshold value TH is experimentally determined as below.

$$TH = \frac{min(\boldsymbol{F(X)}) + max(\boldsymbol{F(X)})}{2} \tag{4}$$

where $\boldsymbol{F(X)} = \{f(x_1), f(x_2), ..., f(x_N)\}$ is a ranking value vector via ranking function f for each $\boldsymbol{A^{inp}}$. $\boldsymbol{X} = \{x_1, x_2, ..., x_N\}$ represents image features for $\boldsymbol{A^{inp}}$. TH is calculated only with the initial masked region in the first iteration and it continues to be used in the subsequent loops. Note that the output value of the ranking function does not represent the absolute score of the inpainting quality. However, the relative relationships of the output values reflect these quality orders. Therefore, we can not set the threshold value TH beforehand, and TH should be determined with relative relationships of the samples.

With the TH, masked super pixels to be reformed $\boldsymbol{S^{ref}} = \{S_1^{ref}, S_2^{ref}, ..., S_M^{ref}\}$ are obtained among $\boldsymbol{S^{mask}}$ by finding corresponding outlier adjacent super pixels with lower ranking value than TH. If $\boldsymbol{S^{ref}}$ exist, the masked region is reformed; i.e., dilated or eroded. Super pixels with $f(x) < TH$ indicate more unnaturalness because we define that positive ranking values are better as shown in Eq. 1.

C. MASKED REGION REFORMATION

This subsection describes algorithms for masked region reformation. The key idea here is that masked regions are

dilated or eroded so that unnatural super pixels do not form the contours of a masked region. This reforming makes it possible to avoid generating unnatural inpainting. Reforming towards dilation or erosion is decided before the optimization. In subsections III-C.1 and III-C.2 we respectively show masked region dilation and erosion algorithms.

1) MASKED REGION DILATION

The basic idea for masked region dilation is that masked regions are iteratively expanded until there are no unnatural adjacent super pixels included in $\boldsymbol{A}^{\boldsymbol{inp}}$. For expansion, a neighbor super pixel of a super pixel to be reformed $\boldsymbol{S}^{\boldsymbol{ref}}$ is added into a masked region. Algorithm 2 and Figure 6 show the pseudo code and figures of this processing.

Algorithm 2 Masked Region Dilation

Input: Masked image I_m, super pixels to be reformed $\boldsymbol{S}^{\boldsymbol{ref}}$
Output: Dilated masked image I_m

```
for n = 1 to |S^ref| do
    S^nei ⇐ neighbor super pixels of S_n^ref
    for l = 1 to |S^nei| do
        for Each pixels p consists of S^nei do
            Retrieve the nearest masked super pixel S_n'^mask
            if S_n'^mask == S_n^ref then
                Add p to masked region in I_m
            end if
        end for
    end for
end for
```

Let $\boldsymbol{S}_{\boldsymbol{n}}^{\boldsymbol{nei}} = \{S_{(n,1)}^{nei}, S_{(n,2)}^{nei}, ..., S_{(n,l)}^{nei}\}$ be neighbor super pixels of S_n^{ref}. S_n^{ref} represents the $n-th$ super pixel to be reformed (See Figure 6(a)). Here, including all the pixels in $\boldsymbol{S}^{\boldsymbol{nei}}$ in the masked region expands the region more than necessary. To avoid this over masking, the method selects pixels to be added into the masked region. This is because in general, smaller masked regions are better unless they do not generate unnatural inpainted regions. For each pixel p consists of $\boldsymbol{S}^{\boldsymbol{nei}}$, the closest masked super pixel is found from $\boldsymbol{S}^{\boldsymbol{mask}}$. Let the center of each S_n^{mask} be O_n as shown in Figure 6(b). For all pixels p, the index of the closest masked super pixel $S_{n'}^{mask}$, i.e., n' is calculated as below.

$$n' = \arg\min_{n} (\text{distance}(p, O_n)), \tag{5}$$

where the function *distance* calculates euclidian distance between two points.

If obtained super pixel $S_{n'}^{mask}$ is equal to S_n^{ref}, p is added to the masked region as shown in the blue region of Figure 6(b). After this processing for each $\boldsymbol{S}^{\boldsymbol{ref}}$, the updated masked region is obtained as shown in Figure 6(c).

2) MASKED REGION EROSION

The basic idea of masked region erosion is iterative removal of masked pixels, until there are no unnatural adjacent super pixels found from $\boldsymbol{A}^{\boldsymbol{inp}}$. For the removal, super pixels to be reformed $\boldsymbol{S}^{\boldsymbol{ref}}$ are excluded while unwanted objects are not revealed. Algorithm 3 and Figure 7 show the pseudo code and figure of this processing.

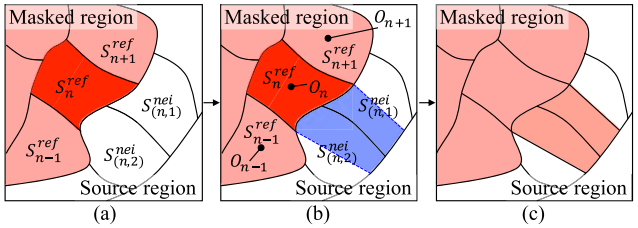


FIGURE 6. Masked region dilation processes. (a) Unnatural masked super pixel to be reformed $\boldsymbol{S}^{\boldsymbol{ref}}$ and its neighbor non-masked super pixels $\boldsymbol{S}^{\boldsymbol{nei}}$ are obtained. (b) To avoid over masking, not $\boldsymbol{S}^{\boldsymbol{nei}}$ as a whole but only the pixels whose nearest masked super pixel is $\boldsymbol{S}^{\boldsymbol{ref}}$ (shown in blue) are added to the mask. (c) Dilated masked region.

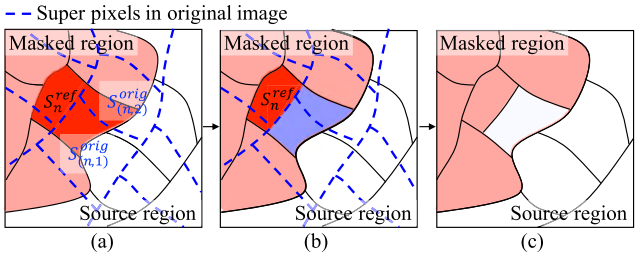


FIGURE 7. Masked region erosion processes. (a) Unnatural masked super pixel to be reformed $\boldsymbol{S}^{\boldsymbol{ref}}$, and super pixels with original image (shown with blue dotted line) that are locationally overlapped with both $\boldsymbol{S}^{\boldsymbol{ref}}$ and non-masked region obtained as $\boldsymbol{S}^{\boldsymbol{orig}}$. (b) To avoid unnecessary exclusion of masks, not $\boldsymbol{S}^{\boldsymbol{ref}}$ as a whole but only pixels overlapped with $\boldsymbol{S}^{\boldsymbol{orig}}$ shown in blue are excluded from masks. (c) Eroded masked region.

Unlike the masked dilation algorithm, super pixels consisting of an original image are also considered. Let $\boldsymbol{S}^{\boldsymbol{orig}} = \{S_{(n,1)}^{orig}, S_{(n,2)}^{orig}, ..., S_{(n,l)}^{orig}\}$ be super pixels generated with original image I_o overlapped with S_n^{ref}. In Figure 7(a), $\boldsymbol{S}^{\boldsymbol{orig}}$ are overlapped onto super pixels in masked images shown in blue dotted lines.

$\boldsymbol{S}^{\boldsymbol{orig}}$ is used for deciding whether pixels are excluded from a masked region. Excluding $\boldsymbol{S}^{\boldsymbol{ref}}$ as a whole may reveal unwanted objects behind the masked region. Since $\boldsymbol{S}^{\boldsymbol{orig}}$ is considered to be suitable for objects in an original image, using $\boldsymbol{S}^{\boldsymbol{orig}}$ can avoid unnecessary exclusion. Each pixel p of $\boldsymbol{S}^{\boldsymbol{ref}}$ is excluded only when $\boldsymbol{S}^{\boldsymbol{orig}}$ where p overlaps the outermost side face (See the region masked with blue in Figure 7(b)). After this erosion for each $\boldsymbol{S}^{\boldsymbol{ref}}$, the updated masked region is obtained as shown in Figure 7(c).

IV. EXPERIMENTS

This section investigates the effectiveness of our proposed method. We will start with the experimental setups to obtain ranking function f in subsection IV-A. Then subsection IV-B investigates the proposed method's effectiveness for finding unnatural inpainted images, compared with other metrics. Subsection IV-C shows the effectiveness of our masked region reformation framework with various

Algorithm 3 Masked Region Erosion

Input: Masked and original images I_m, I_o and super pixels to be reformed $\boldsymbol{S}^{ref}$

Output: reformed masked image I_m

$\boldsymbol{S}^{orig} \Leftarrow$ SuperPixelSegmentation(I_o)
for $k = 1$ to $|\boldsymbol{S}^{ref}|$ **do**
  for $l = 1$ to $|\boldsymbol{S}^{orig}|$ **do**
    if S_k^{ref} and S_l^{orig} are overlapped **then**
      Add S_l^{orig} to $\boldsymbol{S}^{orig}$
    end if
  end for
  for pixels $p = (p.x, p.y)$ consists of $\boldsymbol{S}^{orig}$ **do**
    if p is included in S_k^{ref} **then**
      $I_m(p.x, p.y) \Leftarrow I_o(p.x, p.y)$
    end if
  end for
end for

images and IV-D subjectively evaluates the inpainted results.

A. EXPERIMENTAL SETUP

To generate a training set for ranking function f, 111 images with manually masked unwanted regions were prepared. The 111 images were inpainted with two existing inpainting methods [4], [5]. Six parameter sets (= 3 patch sizes × 2 levels of multi-scale parameters) were used for both methods. We randomly displayed a pair of inpainted images side-by-side as shown in Figure 8. Subjects were asked to choose one of three options: **r**: right image is more natural, **l**: left image is more natural, and **n**: no preference order (i.e., it is hard to decide which one is more natural). Excluding inpainted images with an extremely low level of naturalness and images that did not get a consistent response from all subjects, we prepared 2,466 image pairs.

We implemented RankingSVM with SVM Rank [44] 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) for training f. Eight subjects (four males and four females) with normal vision evaluated which images were more natural. Ranking function f was trained depending on this annotation. The trained f is used for unnatural region detection in next subsection.

B. COMPARISON WITH EXISTING METRICS FOR ESTIMATING UNNATURALNESS

This subsection investigates the effectiveness of our proposed unnatural region estimation for inpainted images using the trained ranking function. We compare our technique with existing IQA methods using computational saliency maps by Zhang *et al.* [42] and Herbrich *et al.* [43] that were used in the IQA methods of Trung *et al.* [26] and Venkatesh and Cheung *et al.* [23].

FIGURE 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.

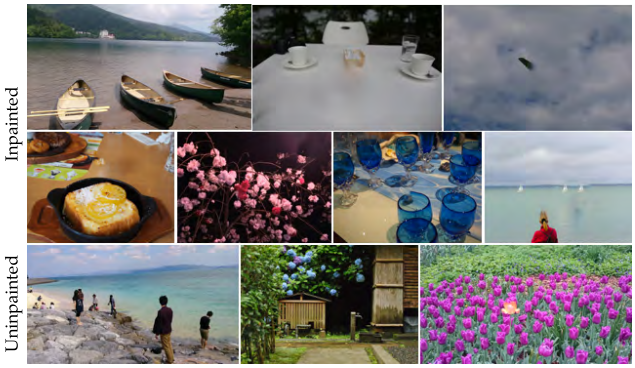


FIGURE 9. Stimulus images for subjective mask annotation. Top and middle rows show inpainted images. Not inpainted images are shown in the bottom row.

FIGURE 10. User interface for pointing out unnatural region in inpainted image.

To prepare ground-truth data, we asked 12 subjects(11 males, one female) with normal vision to draw unnatural regions in 10 images consisting of seven inpainted and three not inpainted images as shown in Figure 9. The latter are original images to which no image processing was applied.

A drawing interface is shown in Figure 10. Subjects were asked to point unnatural regions out in images without any time limitation. They used a mouse as a drawing device and could change the pen size for drawing as they liked. Depending on how hard the subjects pressed the pen, the opacity (brush depth) of the line was changed. Then a heat map was generated from a drawn mask and overlaid on an image as shown in Figure 10. Subjects were informed that the observed images included both inpainted and unpainted images, but

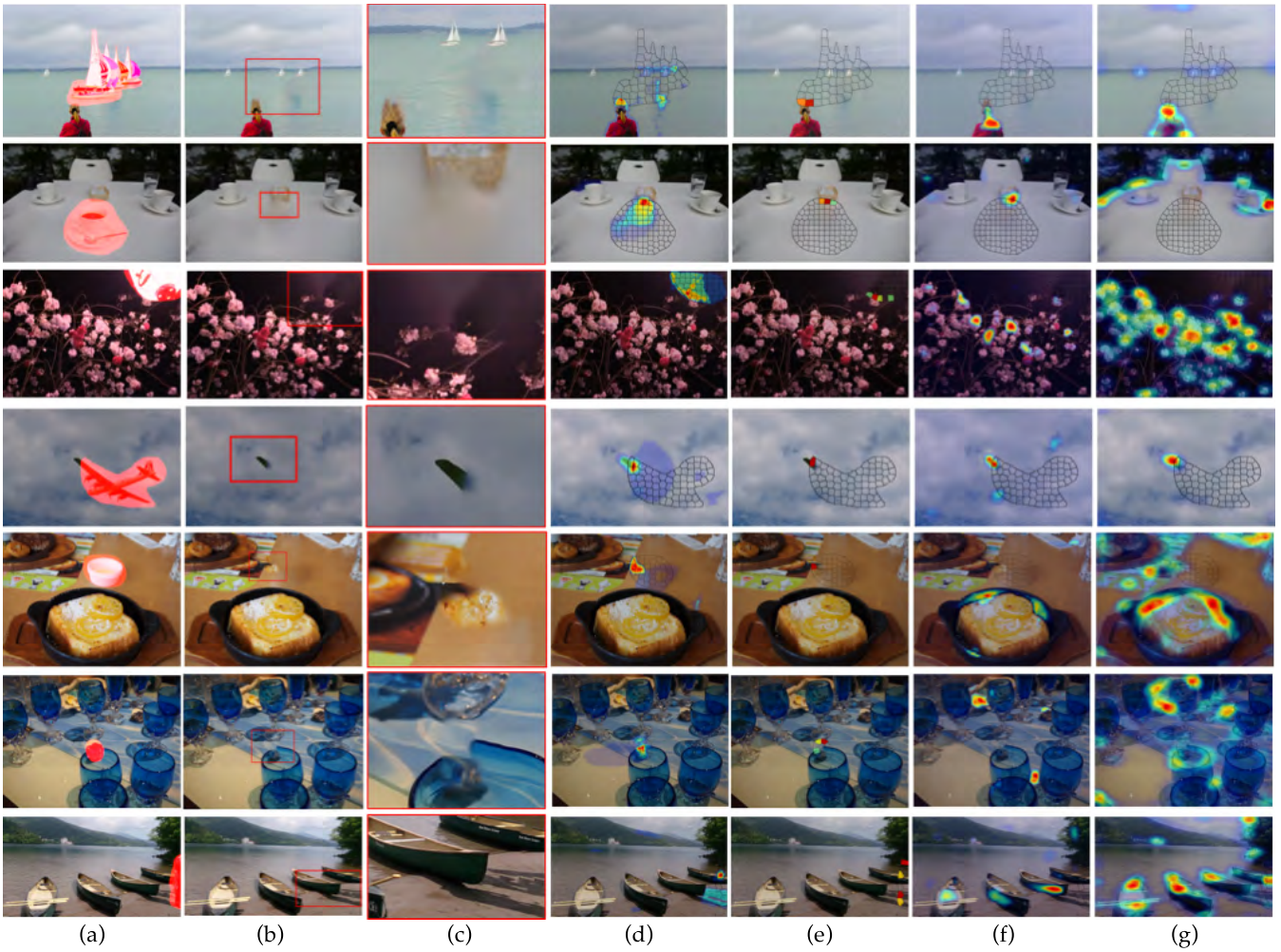


FIGURE 11. Comparison between proposed unnatural region detection and existing metrics for finding unnaturalness with subjectively annotated unnatural region as ground truth. (a) original image with damaged region masked in red, (b) and (c) inpainted image and their close-up views of unnatural regions, (d) subjectively annotated unnatural region. (e)-(g) Obtained unnatural regions with heat maps overlaid on (b) (red gathers more unnaturalness). (e) with proposed method by super pixel basis, (f) with saliency maps by Hou et al's method [42] used in Voronin et al's metric [26], (g) Walther *et al.*'s [43] used in Oncu *et al.*'s metric [23].

were not informed how many images were inpainted in order to prevent them from being affected by prior knowledge. The LCD monitor used for stimulus presentation was 21 inches (1280 × 1080 pixels). The distance between the monitor and the observers was 60 cm.

The inpainted image results are shown in Figure 11. Original images with a region masked in red are shown in (a), (b) and (c) show inpainted images and close-up views of their unnatural regions, (d) shows subjectively annotated unnatural regions as ground-truth, and (e)-(g) show computed unnatural regions as a heat map overlaid on (b) (red gathers more unnaturalness). (e) shows results obtained with the proposed method on a super pixel basis and (f) shows those obtained with the saliency map described by Zhang *et al.* [42] and used in Voronin *et al.*'s IQA metric [26]. (g) shows those obtained with the map described by Herbrich *et al.* [43] and used in Oncu *et al.*'s IQA metric [23].

As shown in the fourth row of Figure 11, for a region that is obviously unnatural such as a part of the wing of the airplane left in the uniform sky texture, all methods correctly simulate human attention. However, as shown in (f) and (g), existing saliency maps failed to simulate human attention in other rows. One of the reasons for this is that there are gaps between human gaze patterns and computational saliency maps as Boykov and Jolly *et al.* [16] revealed. For example, in the first and third rows, the red cloth worn by the woman or the red flowers gather more attention with existing metrics as shown in (f) and (g) because saliency maps are typically designed by assuming that warmer colors gather more gazes. In addition, in the second, fifth, and sixth rows, regions with more edges gather more attention, unlike subjectively annotated attention. The reason for this also comes from typical saliency map designs, which estimate more gazes on stronger edges. At the bottom row, existing metrics could not find unnatural inpainting because of the subtle changes in texture or color, while our metric could do so as shown in (f) and (g). As shown in (e), our method successfully estimates subjects' attention for all image stimuli, indicating

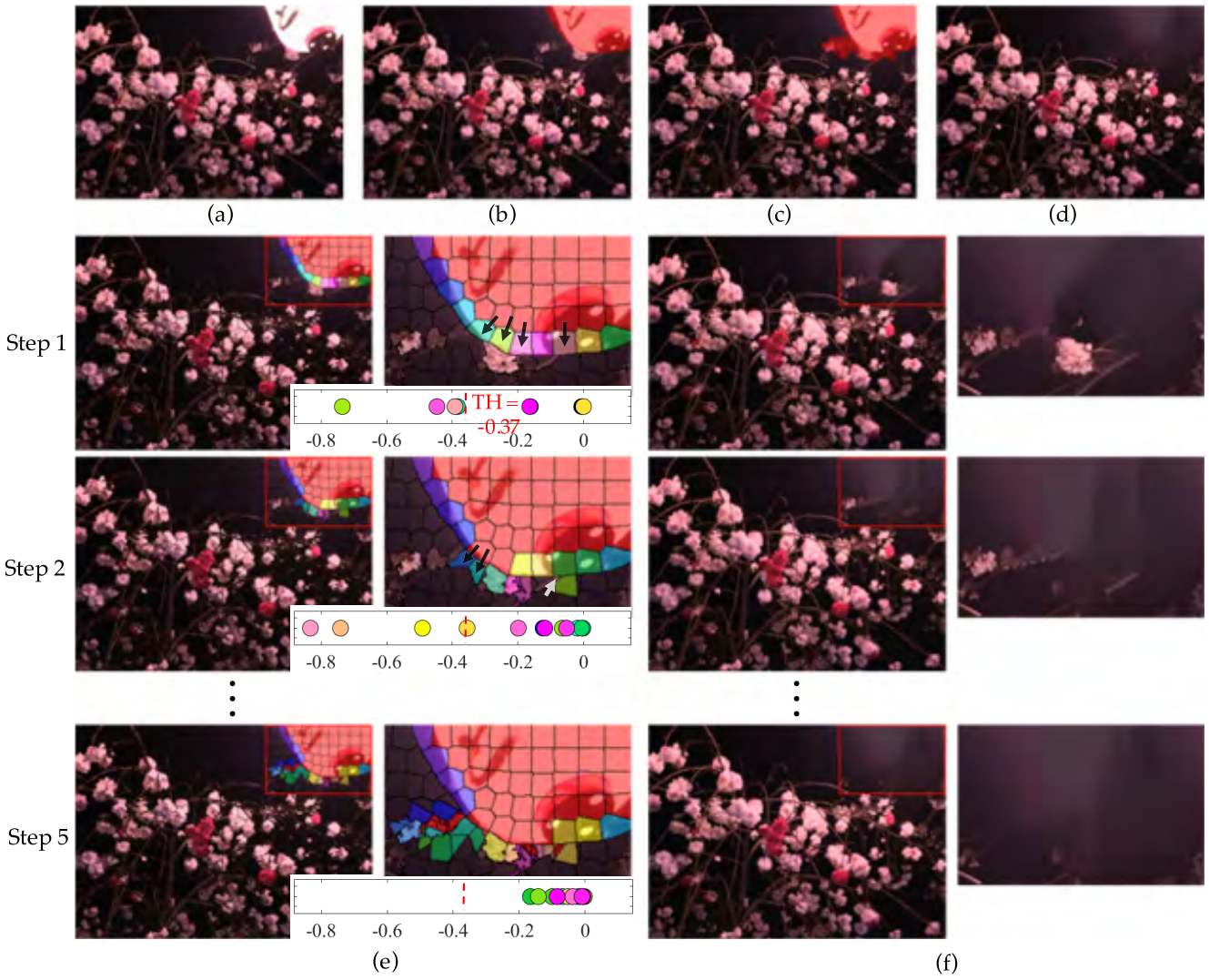


FIGURE 12. Masked region dilation result. (a) original image, (b) initial masked region (shown in red) for (a), (c) final masked image obtained for the reformed masked region, and (d) final inpainted images with (c). Rows (e) and (f) show the output results in each iterative step, and (e) shows the masked image and its close-up view with colored super pixels and their ranking value plots. Colored super pixels excluding red ones are super pixels on the contours of the masked region, which are candidates for dilation. Ranking values for all colored super pixels are also plotted at the bottom right of (e). All plotted samples and super pixels correspond to each other in color. Samples plotted lower than *TH* are outliers that indicate super pixels to be reformed.

that the proposed unnatural region estimation method works effectively.

C. MASKED REGION REFORMATION EXPERIMENTS

This subsection investigates the efficacy of the proposed masked region optimization method. Figures 12 and 13 respectively show masked region dilation or erosion obtained with the proposed method. In both figures, (a) shows an original image, (b) shows an initial masked region (shown in red) for the original image, (c) shows the final masked image obtained for the reformed masked region with the proposed method, and (d) shows the final inpainted images obtained with (c). Rows (e) and (f) show the output results obtained in each iterative step, and (e) shows the masked image and its close-up view with colored super pixels and their ranking value plots. Colored super pixels excluding red ones are super pixels on the contours of the masked region, which are candidates for dilation or erosion. Ranking values via ranking function f for all colored super pixels are also plotted at the bottom right of (e) in each step. All plotted samples and super pixels correspond to each other in color. Samples plotted lower than TH are outliers that indicate super pixels to be reformed. The iterations of unnatural region detection and masked region reformation empirically converge about three to five times.

In Figure 12, the initial masked region shown in (b) hides an unwanted lantern, but is also overlapped with the flowers and branches of a cherry blossom tree. This provides a failed inpainted result that has discontinuities in both color and structure around the flowers or branches (See Step 1 in (f)).

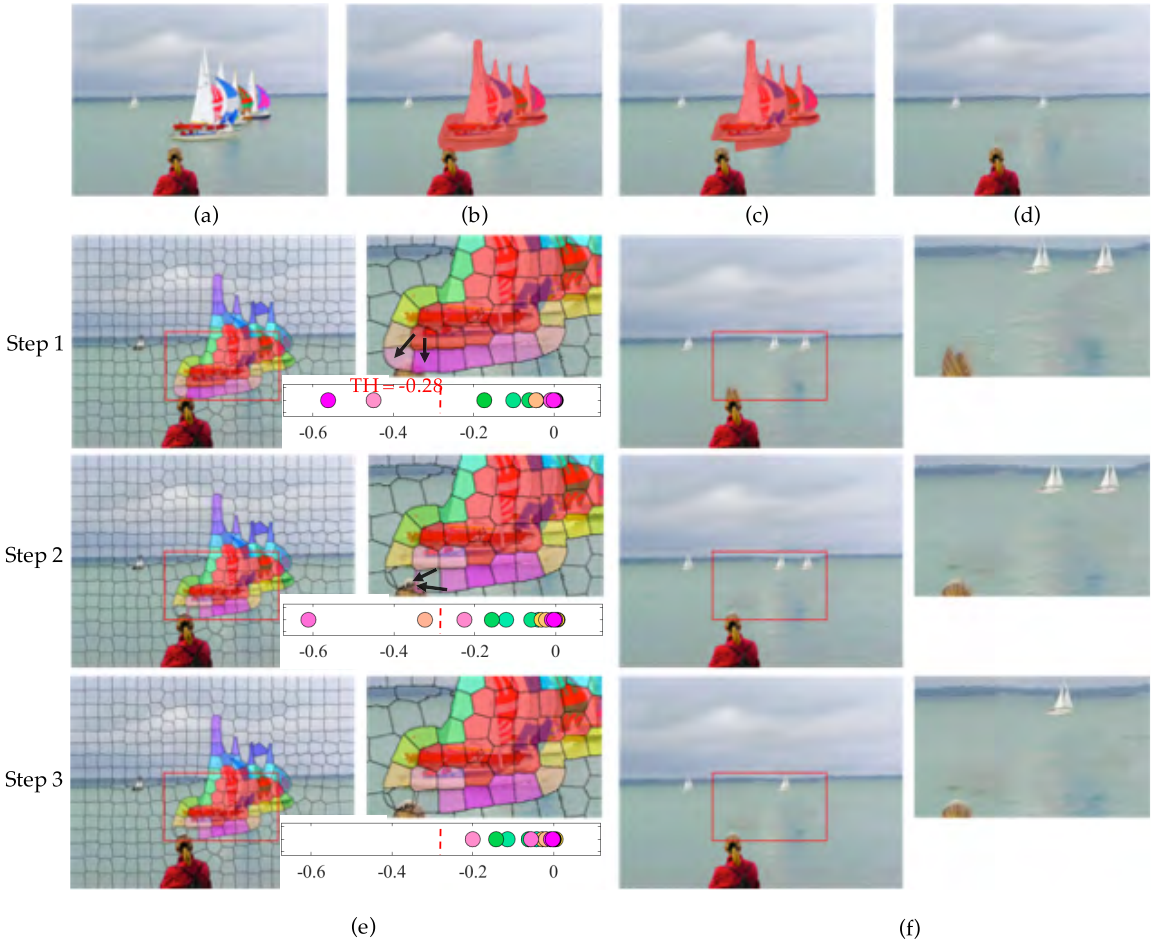


FIGURE 13. Masked region erosion result. (a) original image, (b) initial masked region (shown in red) for (a), (c) final masked image obtained for the reformed masked region, and (d) final inpainted images with (c). Rows (e) and (f) show the output results in each iterative step, and (e) shows the masked image and its close-up view with colored super pixels and their ranking value plots. Colored super pixels excluding red ones are super pixels on the contours of the masked region, which are candidates for erosion. Ranking values for all colored super pixels are also plotted at the bottom right of (e). All plotted samples and super pixels correspond to each other in color. Samples plotted lower than *TH* are outliers that indicate super pixels to be reformed.

To avoid such failures, masked region dilation is performed. All colored super pixels other than those in red in (c) are reformation candidates. In the first iteration step, the threshold to find outlier super pixels was obtained as $TH = -0.37$. In the close-up view in (e), outlier super pixels are annotated with the arrows and masked regions are dilated depending on such super pixels as shown in (e) in the next step. Finally, in step 5, there are no outlier super pixels and good inpainted results are obtained (See step 5 in (f)).

On the other hand, in Figure 13, the initial masked region in (b) masks the yachts seen above the woman. However, the region also masks the woman's head. This provides undesired inpainted results as shown in (f) at step 1, where the woman's head becomes unnaturally larger. In this case, masked region erosion works effectively. In the first iteration step, the threshold to find outlier super pixels was obtained as $TH = -0.28$. Also, in the close-up view in (e), outlier super pixels are annotated with the arrows and masked regions are eroded depending on the super pixels shown in (e) in the next step. In the final step, the masked region excludes the head region of the woman. With this masked region, an inpainted image without any unnatural super pixels is provided.

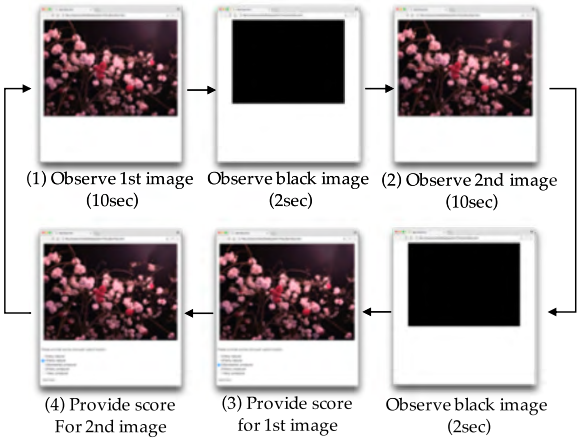


FIGURE 14. Test procedure for providing 5-point scores.

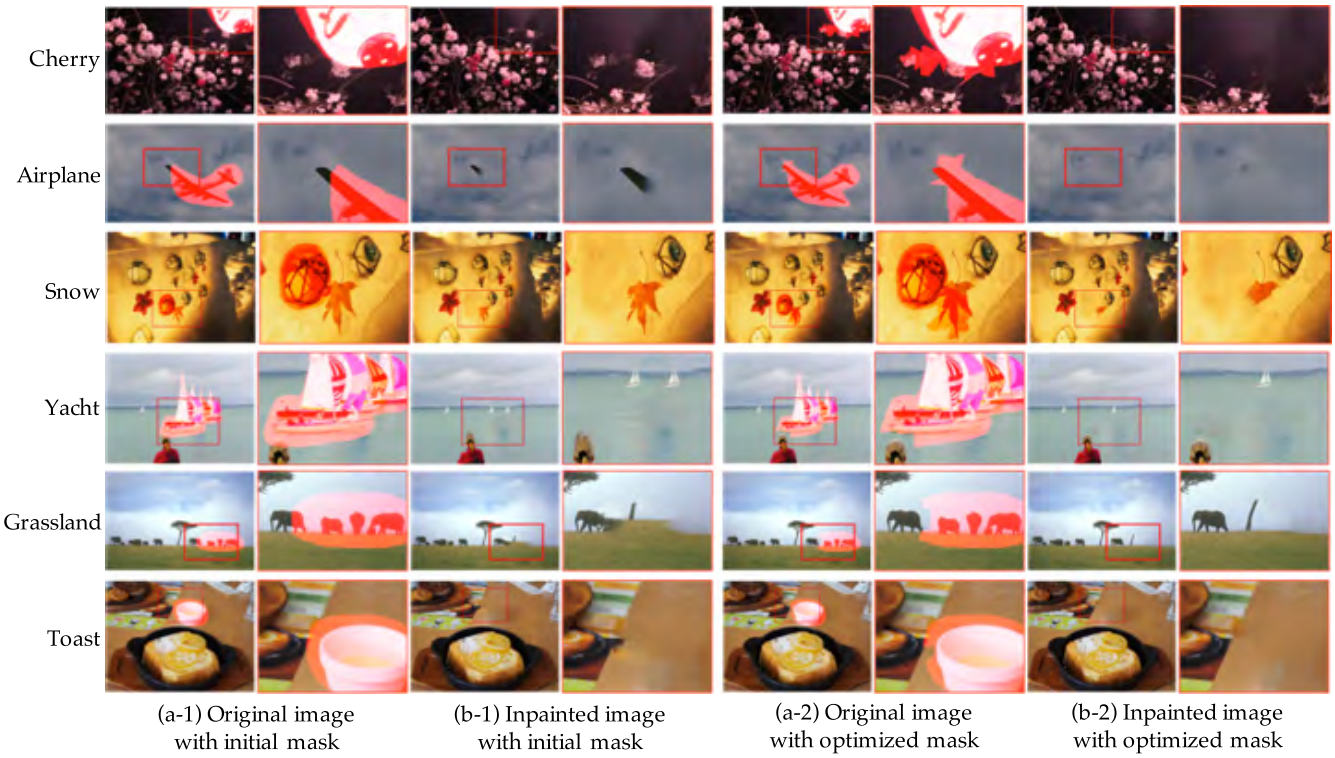


FIGURE 15. Image stimuli for subjective evaluation. (a-1) Original image with initial mask. (b-1) Inpainted image with initial mask. (a-2) Original image with optimized mask. (b-2) Inpainted image with optimized mask.

D. SUBJECTIVE EVALUATION FOR INPAINTED IMAGE QUALITY

This subsection subjectively compares the naturalness of resultant inpainted images depending on initial masked regions and those optimized with the proposed method. Figure 14 shows the test procedure, in which subjects repeated four tasks: (1) observe the first image for 10 seconds, (2) observe the second image for 10 seconds, (3) provide a score for the first image, and (4) provide a score for the second image. We asked 6 subjects (5 males and 1 females) with normal vision to report 5 point scores for each image (1: very unnatural, 2: fairly unnatural, 3: somewhat unnatural, 4: fairly natural, and 5: very natural). Subjects in this experiment did not duplicate those mentioned in IV-B. For the first and second images, inpainted images with initial and optimized masked regions are randomly shown. In order to avoid differences between two images being noticed, uniform black images are shown between tasks (1) and (2).

The image stimuli are shown in Figure 15. In the figure, (a-1) and (b-1) show original images with damaged regions masked in red and their inpainted images with close-up views for initial masked regions, while (a-2) and (b-2) show those for masked regions optimized by the proposed method. The top three rows are for masked region dilation while the bottom three rows are for erosion.

The averaged scores are shown in Figure 16. As shown in the figure, subjective scores for inpainted images with optimized masked regions are improved for all samples except

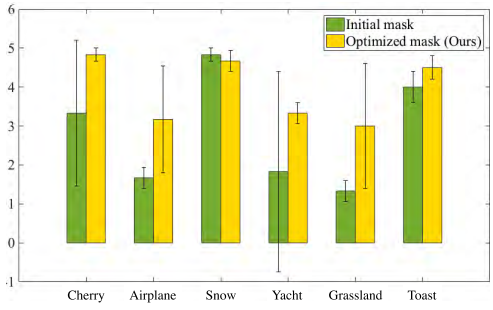


FIGURE 16. Average subjective scores for each images Figure 15.

for "snow." In the first and second rows, the initial masked region overlaps the pink flower and the branch, or the wing of the airplane. These overlaps generate texture discontinuities in the inpainted region. However, because our optimized mask includes such objects, these discontinuities are removed. In the fourth, fifth, and sixth rows, the original masked region overlaps the woman's head, the elephant's body, and the sheet on the table. These masked regions cause unnatural inpainted results due to the difficulty in finding appropriate source regions to fill in the holes. However, our optimized masked regions achieve better results by excluding such object regions from masks. In "snow", the initial masked region overlaps a red leaf and the inpainted image has discontinuity around the leaf. In this case masked region dilation was converged before the mask covers the entire leaf.

This is because the edge of the texture in the inpainted region is changed gradually and generates less unnaturalness. In fact, some subjects answered that the leaves in (b-2) seemed to be buried under the snow and thus it was a natural scene.

V. DISCUSSION

This section briefly reviews the experiments covered in the previous section. Unlike existing IQA methods, our method can appropriately find regions including unnatural areas in inpainted regions. This was shown through a comparison between unnatural regions that our method found and those that subjects drew. As a result of this unnatural region detection, our method effectively reformed masked regions and achieved better inpainted results. Even when inpainted results include unnatural areas, our method excludes them by dilating or eroding initial masked regions.

Here we will also mention our method's limitation. Our learning-to-rank-based unnatural super pixel detection technique depends on color and texture discontinues inside and outside damaged regions. Thus, as shown in "snow" in Figure 15, our natural region detection does not work well for images that are inpainted with blurred colors or textures. One possible improvement to the method is enabling it to take semantic information of unwanted objects into account.

Currently our framework outputs both dilated and eroded masked regions. This is because we cannot determine which generates better results. We believe an acceptable procedure is for users to choose one of them as a last step of the framework. However, we think the procedure in which users choose one of them as a last step of the framework is acceptable.

VI. CONCLUSION

This paper proposed a masked region optimization framework for image inpainting. This is the first method that trials showed automatically erodes or dilates masked regions to be inpainted to achieve good inpainted results. The method also significantly reduces users' working time and the inputs they must provide because it only requires a first input of a masked region. By focusing on a learning-to-rank-based approach to estimate where unnatural inpainted results are generated in masked regions, the proposed method reforms masked regions to ease inpainting tasks. Experimental results showed that this framework effectively works.

Since our framework outputs both dilated and eroded masked regions, an interesting subject for future work will be to introduce another indicator to determine which to choose before the iterative reformation process.

APPENDIX

IMAGE FEATURES FOR LEARNING TO RANK

As image features x dedicated for evaluating inpainted images, we used the 10-dimensional vector $\boldsymbol{x} = (\boldsymbol{X_d}, \boldsymbol{X_s})$, where $\boldsymbol{X_d}$ and $\boldsymbol{X_s}$ represent unnaturalness produced by color or structural discontinuity in an image. All of $\boldsymbol{X_d}$ and $\boldsymbol{X_s}$ have 5-dimensional values. $\boldsymbol{X_d}$ and $\boldsymbol{X_s}$ are computed as below;

$$X_d = ||S(P_{in}) - S(P_{out})||_2^2 \tag{6}$$

$$X_s = \frac{\sum_{p\in\delta\Omega} S(P_{out}(p))}{\sum_{p\in\delta\Omega} 1} \tag{7}$$

where Ω and $\delta\Omega$ respectively denote a masked region and its contour. Eq. 6 represents a squared 2-norm. $P_{in}(p)$ and $P_{out}(p)$ show masked and source regions in patch $P(p)$, which is centered at point p. 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.

$$S(P_{in}(p)) = \frac{\sum_{q\in P(p)\cap\Omega} \boldsymbol{s(q)}}{\sum_{q\in P(p)\cap\Omega} 1} \tag{8}$$

$$S(P_{out}(p)) = \frac{\sum_{q\in P(p)\cap\bar{\Omega}} \boldsymbol{s(q)}}{\sum_{q\in P(p)\cap\bar{\Omega}} 1} \tag{9}$$

In the work we report in this paper, we used $\boldsymbol{s(p)} = (u(p), v(p))$, where $\boldsymbol{u(p)} = (u_R(p), u_G(p), u_B(p))$ and $\boldsymbol{v(p)} = \boldsymbol{v_x(p), v_y(p)}$, each denoting RGB pixel values and two-dimensional edge texture features.

REFERENCES

[1] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 24.

[3] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 82:1–82:10, 2012.

[4] K. He and J. Sun, "Image completion approaches using the statistics of similar patches," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2423–2435, Dec. 2014.

[5] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 129.

[6] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. 25th Adv. Neural Inf. Process. Syst.*, 2012, pp. 341–349. [Online]. Available: https://nips.cc/Conferences/2012/Program/event.php?ID=3279

[7] R. Köhler, C. Schuler, B. Schölkopf, and S. Harmeling, "Mask-specific inpainting with deep neural networks," in *Pattern Recognition*, X. Jiang, J. Hornegger, and R. Koch, Eds. Cham, Switzerland: Springer, 2014, pp. 523–534.

[8] J. S. J. Ren, L. Xu, Q. Yan, and W. Sun, "Shepard convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2015, pp. 901–909. [Online]. Available: http://www.deeplearning.cc/shepardcnn/

[9] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 107.

[10] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4076–4084.

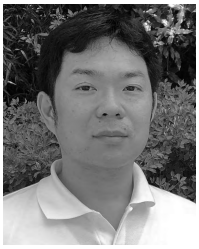
[11] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," in *Proc. ACM SIGGRAPH*, 2004, pp. 309–314.

[12] Q. Yang, C. Wang, X. Tang, M. Chen, and Z. Ye, "Progressive cut: An image cutout algorithm that models user intentions," *IEEE Multimedia*, vol. 14, no. 3, pp. 56–66, Jul./Sep. 2007.

[13] E. N. Mortensen and W. A. Barrett, "Toboggan-based intelligent scissors with a four-parameter edge model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 1999, pp. 452–458.

[14] M. S. Farid, A. Mahmood, and M. Grangetto, "Image de-fencing framework with hybrid inpainting algorithm," *Signal Image Video Process.*, vol. 10, no. 7, pp. 1193–1201, Oct. 2016.
[15] M. S. Farid, M. Lucenteforte, and M. Grangetto, "DOST: A distributed object segmentation tool," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 20839–20862, 2018.
[16] M. Isogawa, D. Mikami, K. Takahashi, and A. Kojima, "Eye gaze analysis and learning-to-rank to obtain the most preferred result in image inpainting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3538–3542.
[17] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2003, pp. 10–17.
[18] C. Guillemot and O. Le Meur, "Image inpainting : Overview and recent advances," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 127–144, Jan. 2014.
[19] T. Shiratori, Y. Matsushita, X. Tang, and S. B. Kang, "Video completion by motion field transfer," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 411–418.
[20] M. Isogawa, D. Mikami, K. Takahashi, and A. Kojima, "Image and video completion via feature reduction and compensation," *Multimedia Tools Appl.*, vol. 76, no. 7, pp. 9443–9462, 2017.
[21] M. V. Venkatesh and S.-C. S. Cheung, "Eye tracking based perceptual image inpainting quality analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1109–1112.
[22] P. A. Ardis and A. Singhal, "Visual salience metrics for image inpainting," *Proc. SPIE*, vol. 7257, pp. 72571W-1–7271W-9, Jan. 2009.
[23] A. I. Oncu, F. Deger, and J. Y. Hardeberg, "Evaluation of digital inpainting quality in the context of artwork restoration," in *Proc. Eur. Conf. Comput. Vis. Workshops Demonstrations (ECCV)*, vol. 7583, 2012, pp. 561–570.
[24] A. D. T. Trung, B. A. Beghdadi, and C. C. Larabi, "Perceptual quality assessment for color image inpainting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 398–402.
[25] V. A. Frantc, V. V. Voronin, V. I. Marchuk, A. I. Sherstobitov, S. Agaian, and K. Egiazarian, "Machine learning approach for objective inpainting quality assessment," *Proc. SPIE*, vol. 9120, pp. 91200S-1–91200S-9, May 2014.
[26] V. V. Voronin, V. A. Frantc, V. I. Marchuk, A. I. Sherstobitov, and K. Egiazarian, "No-reference visual quality assessment for image inpainting," *Proc. SPIE*, vol. 9399, pp. 93990U-1–93990U-8, Mar. 2015.
[27] O. Le Meur, "Robustness and repeatability of saliency models subjected to visual degradations," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 3285–3288.
[28] C. Kim and P. Milanfar, "Visual saliency in noisy images," *J. Vis.*, vol. 13, no. 4, p. 5, 2013.
[29] K.-Y. Chang and C.-S. Chen, "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 785–798, Mar. 2015.
[30] J. Yan, S. Lin, S. B. Kang, and X. Tang, "A learning-to-rank approach for image color enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2987–2994.
[31] T. Abe, T. Okatani, and K. Deguchi, "Recognizing surface qualities from natural images based on learning to rank," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 3712–3715.
[32] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 296–304.
[33] P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in *Proc. Forum Bildverarbeitung*, 2012, pp. 1–12. [Online]. Available: http://www.tu-chemnitz.de/etit/proaut/rsrc/neubert_protzel_superpixel.pdf
[34] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels," EPFL Tech. Rep. 149300, 2010.
[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[36] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.
[37] A. Lucchi, K. Smith, R. Achanta, V. Lepetit, and P. Fua, "A fully automated approach to segmentation of irregularly shaped cellular structures in EM images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2010, pp. 463–471.
[38] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV))*, Nov. 2011, pp. 1323–1330.
[39] F. Yang, H. Lu, and M.-H. Yang, "Robust superpixel tracking," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1639–1651, Apr. 2014.
[40] G. Zhang, X. Jia, and J. Hu, "Superpixel-based graphical model for remote sensing image mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5861–5871, Nov. 2015.
[41] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large-Margin Classifiers*. Cambridge, MA, USA: MIT Press, 2000, pp. 115–132.
[42] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2007, pp. 1–8.
[43] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.
[44] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.

MARIKO ISOGAWA received the B.S. and M.S. degrees from Osaka University, Japan, in 2011 and 2013, respectively. She is currently pursuing the Ph.D. degree in engineering sciences with Osaka University. She is a Researcher with NTT Media Intelligence Laboratories. Her research interests include computer vision, multimedia content handling, and information technologies for enhancing sport performance.



DAN MIKAMI (M'11) received the B.E. and M.E. degrees from Keio University, Kanagawa, Japan, in 2000 and 2002, respectively, and the Ph.D. degree from Tsukuba University in 2012. Since 2002, he has been with Nippon Telegraph and Telephone Corporation. His current research activities are mainly focused on computer vision and information technologies for enhancing sport performance. He was a recipient of the Meeting on Image Recognition and Understanding 2009, the Excellent Paper Award 2009, the IEICE Best Paper Award 2010, the IEICE KIYASU-Zen'iti Award 2010, and the IPSJ SIG-CDS Excellent Paper Award 2013.

DAISUKE IWAI (M'16) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 2003, 2005, and 2007, respectively. He was a Visiting Scientist at Bauhaus-University Weimar, Germany, from 2007 to 2008, and a visiting Associate Professor at ETH, Switzerland, in 2011. He is currently an Associate Professor with the Graduate School of Engineering Science, Osaka University. His research interests include spatial augmented reality and projector-camera systems.

HIDEAKI KIMATA received the B.E. and M.E. degrees in applied physics, and the Ph.D. degree in electrical engineering respectively from Nagoya University, Nagoya, Japan, in 1993, 1995, and 2006. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995, and has been involved in the research and development of video processing of coding, realistic communication, computer vision, and recognition based on machine learning (deep learning). He is currently a Senior Research Engineer and also the Supervisor at NTT Media Intelligence Laboratories. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan, and the Chief Examiner of the Special Interest Group on audio visual and multimedia information processing of the Information Processing Society of Japan.

KOSUKE SATO (M'88) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 1983, 1985, and 1988, respectively. He was a Visiting Scientist at the Robotics Institute, Carnegie Mellon University, from 1988 to 1990. He is currently a Professor at the Graduate School of Engineering Science, Osaka University. His research interests include image sensing, virtual reality, and human interface.