Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Salient Object Detection with Importance Degree

YO UMEKI[1], ISANA FUNAHASHI[2], TAICHI YOSHIDA[2](Member, IEEE), and MASAHIRO IWAHASHI[1] (Senior Member, IEEE)

[1]Department of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology, Nagaoka, Niigata, 940-2137 Japan(e-mail: umeki@stn.nagaokaut.ac.jp, iwahashi@vos.nagaokaut.ac.jp)

[2]Department of Communication Engineering and Informatics, University of Electro-Communications, Chofu-shi, 182-8585 Japan(e-mail: i.funahashi@uec.ac.jp, t-yoshida@uec.ac.jp)

We thank Irina Entin, M. Eng., and Maxine Garcia, Ph.D., from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript. This work was supported by JSPS KAKENHI under Grant 16K18104.

ABSTRACT In this paper, we introduce salient object detection with importance degree (SOD-ID), which is a generalized technique for salient object detection (SOD), and propose an SOD-ID method. We define SOD-ID as a technique that detects salient objects and estimates their importance degree values. Hence, it is more effective for some image applications than SOD, which is shown via examples. The definition, evaluation procedure, and data collection for SOD-ID are introduced and discussed, and we propose its evaluation metric and data preparation, whose validity is discussed with the simulation results. Moreover, we propose an SOD-ID method, which consists of three technical blocks: instance segmentation, saliency detection, and importance degree estimation. The saliency detection block is proposed based on a convolutional neural network using the results of the instance segmentation block. The importance degree estimation block is achieved using the results of the other blocks. The proposed method accurately suppresses inaccurate saliencies and estimates the importance degree for multi-object images. In the simulations, the proposed method outperformed state-of-the-art methods with respect to the F-measure for SOD; and Spearman's and Kendall rank correlation coefficients, and the proposed metric for SOD-ID.

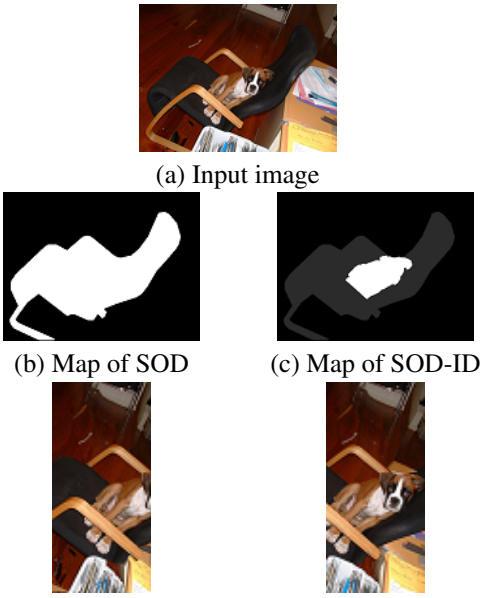
INDEX TERMS Saliency detection, salient object detection, instance segmentation, convolutional neural network (CNN), rank correlation metric

I. INTRODUCTION

SALIENCY detection (SD) is an image processing technique that estimates salient local regions in images [1]–[7]. Salient regions are generally defined as areas that attract human attention with respect to characteristics such as high contrast, unique orientation, and distinctive color. Detecting these regions is important for image applications, such as human eye fixation estimation and context-aware image coding.

Recently, several methods have been proposed for salient object detection (SOD) which is similar to SD [8]–[28]. Instead of estimating local regions, SOD identifies characteristic objects, such as a tall man, a red car, or signs. Some image processing applications require not only salient information but also important object locations [29]–[32]. For example, image retargeting uses the object locations and resizes images while retaining their shapes. Thus, SOD has been shown to be more useful than SD for some applications.

Moreover, Islam *et al.* proposed an expansion of SOD [26], which is called RSOD in this paper, and studies have shown that it has high potential for image applications. SOD classifies estimated objects as salient or non-salient, whereas RSOD estimates salient object contours and their importance scores. Importance scores are useful for several applications, which we show in Fig. 1, where (a) is an input image, (b) and (c) are its ideal saliency map in SOD and RSOD, and (d) and (e) are the retargeting results for (a) using (b) and (c) according to [31], respectively. In (b), the white and black areas represent salient and non-salient regions, respectively, whereas in (c), the white, gray, and black areas represent first salient, second salient, and non-salient regions, respectively. In (d), a part of the dog, which seems to be the most important object in (a), is cropped because the chair and dog are given the same importance value by SOD shown in (b). By contrast, because of the different scores in (c), the dog is completely preserved in (e). Fig. 1 shows one advantage of the expansion, and we experimentally understand that it has high potential



(a) Input image

(b) Map of SOD (c) Map of SOD-ID

(d) Result based on (b) (e) Result based on (c)

FIGURE 1: Retargeting simulation based on SOD and SOD-ID.

not only for image retargeting but also, for example, for content-aware image coding and image representation.

However, the discussion of RSOD was not sufficient in [26] to introduce it as a new theme of computer vision. The authors presented the expansion with little detail as a supplement to the main topic. The details of its definition were omitted, and its significance was not discussed. As the evaluation metric, Spearman's rank correlation coefficient [33] was use simply, but, unfortunately, its validity was not discussed. This inadequate discussion is a problem for tackling the new theme.

In this paper, we call the technique, which is denoted by RSOD, SOD with importance degree (SOD-ID); refine it to introduce a new theme via discussing its definition, significance, and assessment; and propose an SOD-ID method that outperforms state-of-the-art methods. First, we discuss and construct the definition and significance of SOD-ID using several pieces of evidence and application examples. We define the importance score as represented in N degrees, and refer to this as the importance degree. Based on this discussion, we present the evaluation procedure and the dataset preparation of SOD-ID. We also propose an assessment index based on the squared error and Kendall rank correlation coefficient [34], and create a dataset based on those of SD and instance segmentation. Finally, we propose an SOD-ID method based on deep learning and instance segmentation.

Our contributions are summarized as follows:

- We introduce and define a new theme, SOD-ID, via a discussion and examples.
- We introduce the importance degree using N, which is the generalized importance degree of SOD, and show its efficacy and advantages.
- We introduce valid dataset preparation and an effectual evaluation procedure for SOD-ID to evaluate methods without actually applying them to image applications, which contributes to the development of this theme.
- We propose an SOD-ID method via combining instance segmentation and SD based on deep learning as a separable system, which is also useful for SOD.

In simulations, the proposed method perceptually demonstrated N-degree salient objects, and objectively outperformed RSOD [26]. The proposed method accurately detected salient objects and estimated their values of importance degree. The proposed method was objectively compared with RSOD for Spearman's rank correlation coefficient [33], the Kendall rank correlation coefficient [34], and the proposed metric, and obtained better scores. Moreover, in the evaluation procedure of SOD, the results of the proposed method were objectively comparable with state-of-the-art SOD methods; therefore, we demonstrated that it is as effective as SOD.

The remainder of this paper is organized as follows: In Section II, we provide an overview of existing methods of SD, SOD, fixation estimation, semantic segmentation, and instance segmentation. In Section III, we briefly present the fundamentals of SOD and RSOD. In Section IV, we discuss the definition, significance, evaluation metric and dataset of SOD-ID. In Section V, we explain the proposed index, dataset, and method. Finally, we present experimental comparisons in Section VI, and conclude this paper in Section 7.

II. RELATED WORKS

In this section, we explore existing methods of SD, SOD, semantic segmentation, and instance segmentation. First, we describe SD and human eye fixation. Second, we explain two types of SOD and RSOD. Finally, we discuss the difference between semantic segmentation and instance segmentation, and review their recent methods.

SD is similar to human eye fixation; that is, they estimate regions of interest that correspond to human attention [1]–[7], [35], [36]. The traditional SD method uses characteristic features, such as high contrast, unique orientation, and distinctive color [1]. Harel *et al.* proposed a method that uses a graph-based algorithm, calculates activation maps based on several features, and combines them to generate one saliency map [2]. Recently, methods based on a convolutional neural network (CNN) have been proposed, and effectively extract global and complex features as a result of training using a large number of images and their corresponding gaze information [6], [7]. Although they accurately estimate human interest, they cannot estimate object contours.

SOD simultaneously estimates object regions and whether they are salient [8]–[25], [27], [28]. Traditional SOD methods use the propagation algorithm [11], [22], [37]. They iteratively propagate salient and background information based on color similarities between neighboring pixels and the Markov absorption probability. However, they often produce

inaccurate results along object boundaries. Recent methods based on fully convolutional network (FCN) architectures have successfully reduced inaccurate detection. Liu *et al.* proposed a deep hierarchical saliency network that realizes coarse-to-detailed estimation for salient objects [21]. Another method adopts a recurrent network to consider the connection of salient pixels [16].

Although a major SOD dataset contains the importance degree for objects [10], existing methods produce binary results; that is, they classify detected objects into salient or non-salient. The PASCAL-S dataset provides integer saliency values in $[0, 255]$ with object contours. However, SOD methods disregard the priority of each object, and instead focus on estimating the contours of salient objects. Because detecting salient objects and their correct contours is a challenging task, researchers generally propose the estimation of the priority of each detected object as future work.

Semantic segmentation is a technique that identifies categories to which pixels belong, such as human, tree, and car [38]–[40]. Traditional semantic segmentation uses contour detection and the histogram of oriented gradients feature [38]. Recently, the FCN, which is a breakthrough approach for semantic segmentation, has been used to successfully detect image regions [39]. However, semantic segmentation methods cannot separate objects that belong to the same category.

Instance segmentation is derived from semantic segmentation and can identify not only object classes but also their instances [41], [42]. A basic instance segmentation method uses an FCN to detect small windows that each include one object [42]. Another method uses the recurrent architecture to iteratively detect object regions based on previous detection results [41]. Although instance segmentation and SOD similarly detect object contours, instance segmentation disregards their importance; therefore, the purposes of the approaches has been shown to be different.

III. FUNDAMENTALS OF SOD

A. PASCAL-S DATASET

The PASCAL-S dataset contains images, their fixation data, and their SOD maps with multiple values that can be used ground truth (GT) for SOD-ID [10]. It contains 850 natural images whose full segmentation masks are provided in [43]. The fixation data were obtained by applying an eye-tracker to eight subjects that were instructed to perform a free-viewing task for images. In the SOD experiment, 12 subjects were given images and asked to highlight salient objects by clicking on them. The pixels of the SOD maps have integer values in $[0, 12]$, and they are linearly normalized in $[0, 255]$ for the png format. Therefore, we believe that PASCAL-S is an SOD-ID dataset with 13 degrees.

B. FCN METHODS

The FCN was introduced for image classification based on Visual Geometry Group (VGG) networks in [44], and then several FCN architectures were proposed for several applications [6], [35], [36], [39]. The VGG architecture consists of five blocks that each have two or three convolutional layers and a pooling layer. The FCN architecture is constructed by replacing the last layer of the VGG architecture with a one-channel convolutional layer. Some methods that apply merge and convolution layers to the FCN obtain superior results to past methods because the layers realize both shallow and deep convolutions; thereby, they can capture both global and local features [6], [39].

(a) (b)

FIGURE 2: (a) Input image and (b) GT map of SOD-ID.

TABLE 1: Number and percentage of images in the PASCAL-S dataset [10] with respect to the number of salient objects.

♯ Salient objects	1	2	3	4	5	6	7+
♯ Images	300	227	136	72	43	28	44
Distribution (%)	35	27	16	8	5	3	5

C. LOCATION-BIASED DETECTION

In SD and SOD, the location assumption is generally used as prior information [7], [11], [13], [15], [25], [37]. Photographers generally center interesting objects in images, and thus natural images often present salient areas at their center. To exploit this tendency, some SOD methods apply higher weights to salient pixels closer to the center of images [13], [25]. Following this strategy, in an SD method, a location-biased convolution layer was introduced in the FCN, which obtained superior results [7].

D. RSOD

The CNN model detects the contours of salient objects and estimates their multiple saliency values because of its architecture [26]. The architecture recursively calculates saliency maps from coarse to fine levels, and finally fuses the resultant saliency maps. The calculation units are learned using the multi-stage GT of the saliency maps that is generated from PASCAL-S by thresholding its saliency maps at various values. Therefore, the fused maps have various pixel values that reflect saliency levels from coarse to fine.

As an additional process, the method estimates the importance score for each salient object from the output saliency map [26]. In basic terms, the score value is calculated by averaging the saliency values of pixels within the object as

$$Rank(S(X)) = \frac{\sum_{i \in \Omega_X} \chi_i}{N_X}, \tag{1}$$

where S, X, Ω_X, χ_i, and N_X denote a predicted saliency map, candidate salient object, set of indices of pixels that belong to X, saliency value of the i-th pixel, and total number of pixels in X, respectively. It is unknown whether the calculated values are normalized because this is not clearly described in [26]. Note that the authors used the GT segmentation masks in PASCAL-S in this process.

In experiments, the method simply uses conventional methods for evaluation. Spearman's rank correlation coefficient [33] is used as the evaluation metric, and the resultant scores are linearly normalized in $[0, 1]$. PASCAL-S without images used in training is directly used for testing the method.

IV. DISCUSSION ON SOD-ID

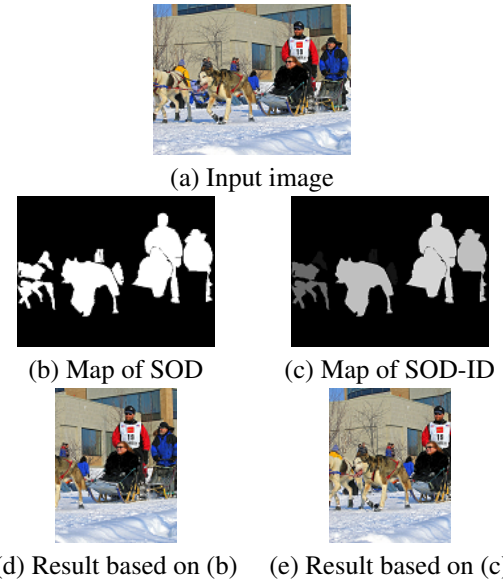
A. DEFINITION OF SOD-ID

We define SOD-ID as a technique that detects the contours of salient objects and estimates their importance degree. Its methods produce a saliency map whose pixel values represent the importance degree scores of objects to which they belong. SOD-ID is mostly similar to SOD, but in contrast to the binary maps of SOD, its GT saliency maps have several values for N-degree objects, as shown in Fig. 2. N-degree means that the maps have integer values in $[0, N-1]$, where, clearly, zero indicates that the pixel of the map belongs to a non-salient object, and N-degree is linearly normalized according to the coding format As mentioned in Section III-A, PASCAL-S seems to be an SOD-ID dataset which $N = 13$ according to experiments. Moreover, note that SOD-ID is a generalized version of SOD; that is, SOD is an SOD-ID in $N = 2$.

In this paper, $N = 7$ is empirically used based on the characteristics of natural images. Table 1 shows the distribution of natural images in PASCAL-S [10] with respect to the number of salient objects within them, where the first, second, and last rows denote the number of salient objects, number of images that include salient objects of the corresponding number in the first row, and distribution, respectively, and "7+" in the eighth column indicates seven or more salient objects. From Table 1, natural images typically contain six or fewer salient objects. They rarely contain seven or more salient objects, but in most cases, some objects in one image have the same saliency levels. Therefore, because 7 degrees adequately realizes SOD-ID for natural images, $N = 7$ is generally valid. Clearly, the value of N can be fixed flexibly for various image applications.

B. SIGNIFICANCE OF SOD-ID

SOD-ID is a generalization of SOD and more suitable for image applications than SOD. People ordinarily rank objects in an image with respect to their interests. Similarly, it has been observed in experiments that subjects sometimes recognize salient objects as non-salient because of the objects' locations. SOD-ID estimates general results of this ranking, and therefore saliency information produced by SOD-ID is more related to human behavior than SOD. Moreover, by thresholding with various parameters as post-processing, SOD-ID produces various saliency maps of SOD. SOD-ID, which is used as pre-processing, results in a variety of saliency information useful to image applications, such as retargeting, content-aware coding, and summarizing.



(a) Input image

(b) Map of SOD

(c) Map of SOD-ID

(d) Result based on (b)

(e) Result based on (c)

FIGURE 3: Retargeting simulation for multi-object images based on SOD and SOD-ID.

For instance, SOD-ID is clearly more suitable than SOD for image retargeting from our experiments. Similar to Fig. 1, Fig. 3 shows retargeting results according to [31] for a multi-object image. The input image in Fig. 3 (a) represents "dogs pull a sled and a human rides" and therefore its important words are "dog," "pull," "sled," "human," and "ride." Image retargeting should retain important words and sentences for input images in its results. In that sense, Fig. 3 (d) shows a failure because the dog is not clearly visible and hence, unfortunately, it represents the wrong sentence, "something pulls a sled and a human rides." By contrast, in Fig. 3 (e), the retargeting result for SOD-ID, accurately represents the original sentence, "dogs pull a sled and a human rides." For other images and retargeting methods, the results sometimes demonstrate the superiority of SOD-ID for image retargeting, as shown in Figs. 1 and 3.

C. SUPERVISED EVALUATION METRIC

The supervised evaluation metric of SOD-ID should measure the degree of similarity with respect to segmentation and the importance degree. Because SOD-ID methods aim to detect the contours of salient objects, they should be evaluated in the same manner as segmentation. Additionally, they should be evaluated when calculating the correlation and similarity of values for scores of the importance degree. An object that has higher scores than another object in the GT should have higher scores in the results of SOD-ID methods, and smaller is better in terms of the difference between the score values between the GT and the results of SOD-ID methods. Unfor-

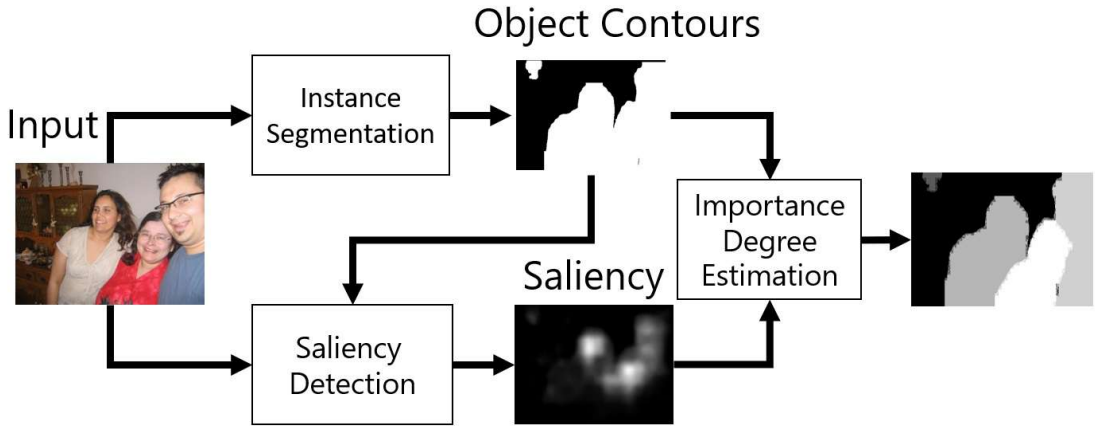


FIGURE 4: Overview of the proposed method.

tunately, because conventional rank correlation coefficients evaluate the correlation but ignore the similarity of scores, Spearman's rank correlation coefficient, which is used in [26], is unsuitable for calculating the importance degree.

In this paper, we propose an evaluation metric for the importance degree of SOD-ID. As the evaluation metric for segmentation, conventional methods, for example, the F-measure, can be used. An evaluation metric for SOD-ID is defined as a linear combination of the F-measure and the proposed metric, or the parallel use of them. The proposed metric F is defined based on simply combining metrics for the correlation and score similarity as

$$F(\boldsymbol{v_p}, \boldsymbol{v_t}) = \alpha R(\boldsymbol{v_p}, \boldsymbol{v_t}) + (1-\alpha) I(\boldsymbol{v_p}, \boldsymbol{v_t}), \quad (2)$$

where R, I, α, $\boldsymbol{v_p}$, and $\boldsymbol{v_t}$ denote the correlation and similarity metrics, a balancing free parameter, and vectors for which each element is the score value of each object, respectively. We use the Kendall rank correlation coefficient as R [34] because it straightforwardly evaluates the correlation and therefore is more suitable than Spearman's rank correlation coefficient. For $\boldsymbol{I}$, we use the squared error and define it as

$$I(\boldsymbol{v_p}, \boldsymbol{v_t}) = \frac{1}{N} \sum_{i=1}^{N} exp(-(v_{pi} - v_{ti})^2 / (2\sigma^2)), \quad (3)$$

where N, v_{pi}, and v_{ti} denote the number of objects, and the i-th element of $\boldsymbol{v}_p$ and $\boldsymbol{v}_t$, respectively, and σ is a free parameter that controls the variance of the Gaussian distribution. R that outputs real values in $[-1, 1]$ is linearly normalized in $[0, 1]$, I has real values in $[0, 1]$ because of (3), and α is restricted in $[0, 1]$, and hence F outputs real values in $[0, 1]$. The metric proposition requires much experimental evidence, but because of the limited space in this paper, the validity of F is briefly shown in Section 6 and a detailed discussion on this topic remains as future work.

TABLE 2: Scores for the estimation methods of the importance degree for the PASCAL-S dataset [10].

	Sum.	Ave.
Kendall score	0.846	0.726
Spearman's score	0.864	0.737

TABLE 3: Construction details of the proposed CNN architecture.

	Name	Size	Stride	Channel
(a)	Conv. 1-1	3×3	1	64
	Conv. 1-2	3×3	1	64
	Pool. 1	2×2	2	64
	Conv. 2-1	3×3	1	128
	Conv. 2-2	3×3	1	128
	Pool. 2	2×2	2	128
	Conv. 3-1	3×3	1	256
	Conv. 3-2	3×3	1	256
	Conv. 3-3	3×3	1	256
	Pool 3	2×2	2	256
	Conv. 4-1	3×3	1	512
	Conv. 4-2	3×3	1	512
	Conv. 4-3	3×3	1	512
	Pool. 4	2×2	1	512
	Conv. 5-1	3×3	1	512
	Conv. 5-2	3×3	1	512
	Conv. 5-3	3×3	1	512
(b)	Conv. 6-1	3×3	1	512
	Conv. 6-2	3×3	1	512
(c)	pPool. 1	2×2	2	512
	pConv. 1	3×3	1	64
	pPool. 2	2×2	4	512
	pConv. 2	3×3	1	64
	pPool. 3	2×2	5	512
	pConv. 3	3×3	1	64
	Conv. 7-1	3×3	1	512
	Conv. 7-2	3×3	1	64
	Conv. 7-3	3×3	1	1

D. DATASET PREPARATION

To create SOD-ID datasets, the procedure of PASCAL-S mentioned in Section III-A is suitable. The segmentation masks are simply obtained manually, and the importance

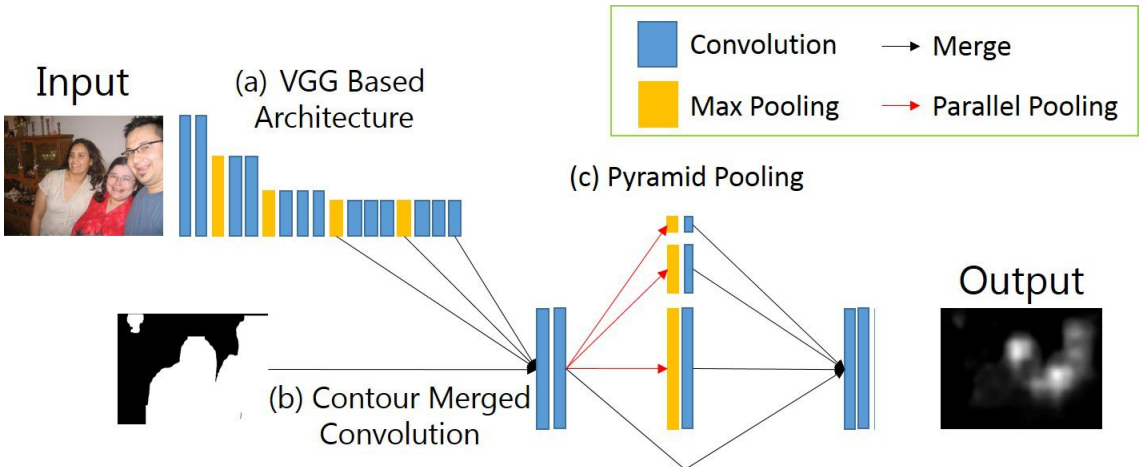


FIGURE 5: Architecture of the proposed CNN method.

TABLE 4: Correlation scores of pairs of arbitrary vectors.

Vectors		Spearman's [33]	Kendall [34]	Prop. metric
[2,6,3]	[1,5,3]	1.000	1.000	0.961
[4,1,6]	[4,2,4]	0.834	0.667	0.748
[6,2,7]	[2,1,3]	1.000	1.000	0.692
[2,1,3]	[1,3,2]	0.250	0.333	0.562

degree is determined as follows: By the strict rules, the subjects of experiments are asked to collect and rank interesting objects in one image. The strict procedure requires several subjects, but unfortunately, it is a difficult task for them. By contrast, the procedure of PASCAL-S only asks subjects to collect interesting objects. For an object, the number of subjects that recognize it as salient is directly determined as its values of the importance degree, and to create a GT map of SOD-ID, pixels within each salient object have their scores based uniformly on the segmentation mask. If M subjects are applied, the resultant map has M degrees. This is simple and useful, but a large number of subjects are required to create general datasets.

To avoid experiments using subjects, we introduce a preparation procedure for the SOD-ID dataset based on existing SD data. As mentioned above, subjective experiments have the troublesome characteristic of requiring many people and large costs. To avoid this, we use existing SD data to produce the SOD-ID maps. The proposed procedure calculates the sum of pixel values within objects in the GT maps of SD, and resultant values are considered as their scores of the importance degree, which is defined in one image as

$$Deg_i = \frac{\sum_{j\in\Omega_i} s_j}{\max_i\{\sum_{j\in\Omega_i} s_j\}}, \tag{4}$$

where Deg_i, s_j, and Ω_i denote the score of the i-th object, i-th pixel value of the SD map, and a set of indices of pixels within the i-th object, respectively. To produce the SOD-ID map, pixel values within the i-th object are uniformly set as Deg_i, and the resultant map is linearly quantized using N. Because the GT maps of SD represent the degree of saliency for each pixel, the summation values within an object are approximately recognized as the degree of interest for the object. Similarly, a pixel value within an object in the GT maps of SD is approximately considered as the number of subjects that recognize the object and categorize it as salient, and therefore, in the case of a large number of subjects, the summation procedure is recognized as the same as that of PASCAL-S for SOD mentioned in Section III-A. Based on the above assumptions, we believe that the proposed procedure is valid for creating SOD-ID datasets.

We experimentally show that the proposed procedure mentioned above has high validity compared with the RSOD procedure mentioned in Section III-D [26]. Using these procedures, SOD-ID maps are produced using the full segmentation masks and fixation data of PASCAL-S. Table 2 shows this comparison, where "Sum." and "Ave." denote the results of the proposed and RSOD procedures; that is, they show values of the evaluation metrics between the SOD maps of PASCAL-S and their resultant maps, respectively. For simplicity, we use Spearman's and Kendall rank correlation coefficients as the metrics [33], [34]. From Table 2, the proposed procedure is clearly better than the RSOD procedure and thus our opinions mentioned above has been shown to be valid.

V. PROPOSED SOD-ID METHOD

A. OVERVIEW

The proposed SOD-ID method is briefly shown in Fig. 4. The system consists of three technical blocks: instance segmentation, SD, and importance degree estimation. First, instance segmentation is applied to an input image to detect object contours, and its arbitrary method can be used here such as that in [41], [42], [48], [49]. Second, the salient regions of the input image are detected by the proposed CNN method using object contours detected in the first block. Finally, using the results of the first and second blocks, the proposed method outputs an SOD-ID map with N degrees through the estimation block of the importance degree. The technical blocks

TABLE 5: F-measure scores of the SOD methods for the DUTS dataset [45].

Image	HDCT [15]	RFCN [16]	DHS [21]	DSSOD [25]	RSOD [26]	Prop.
Image 1	0.120	0.253	**0.745**	0.608	0.000	0.000
Image 2	0.396	0.330	**0.945**	0.933	0.000	0.000
Image 3	0.732	0.794	0.750	0.748	**0.875**	**0.875**
Image 4	0.669	0.614	0.894	0.898	**0.904**	**0.904**
Average	0.521	0.509	**0.761**	0.746	0.711	0.686

TABLE 6: F-measure scores of the SOD methods for the PASCAL-S dataset [10].

Image	HDCT [15]	RFCN [16]	DHS [21]	DSSOD [25]	RSOD [26]	Prop.
Mini bike	0.875	0.704	0.879	**0.911**	0.840	0.840
Car	0.284	0.283	0.597	0.919	**0.960**	**0.960**
Dog	0.575	0.728	0.749	0.917	**0.940**	**0.940**
Horse	0.885	0.699	**0.928**	0.898	**0.928**	**0.928**
Traffic	0.687	0.688	0.772	0.935	**0.965**	**0.965**
Average	0.623	0.584	0.778	0.788	0.793	**0.794**

TABLE 7: F-measure scores of the SOD methods for the SALICON-based dataset [46], [47].

Image	HDCT [15]	RFCN [16]	DHS [21]	DSSOD [25]	RSOD [26]	Prop.
Parking	0.687	0.501	0.766	0.736	0.755	**0.781**
Party	0.565	0.684	0.817	0.736	0.902	**0.937**
Woman	0.696	0.719	0.876	0.731	**0.936**	**0.936**
Bag	0.702	0.753	0.839	0.876	**0.944**	**0.944**
Man	0.528	**0.600**	0.495	0.372	0.419	0.528
Baseball	0.784	0.654	0.842	0.873	0.776	**0.941**
Average	0.526	0.581	0.708	**0.727**		

can be independently developed, and therefore the system provides suitable expandability and serves as a fundamental design of SOD-ID methods.

B. PROPOSED CNN METHOD FOR SD

In this section, we explain the proposed CNN method for SD in the second block that uses the detected contours of the first block. The architecture uses the contours as a part of the input and extracts their multi-resolution features to estimate the saliency values. The loss function imposes different weights for object and background regions based on the contours. Note that the proposed CNN method considers location bias similar to conventional SD and SOD methods.

1) Architecture

Fig. 5 and Table 3 show the architecture of the proposed CNN method and its parameters, respectively. Figs. 5 (a)–(c) correspond to Tables3 (a)–(c), respectively. In Table 3, "Conv.", "Pool.", and "p∗" indicate convolution, max pooling layers, and the pyramid pooling module, respectively. The rectified linear unit [50] is used as the activation function in the convolution layers. A VGG-based method is used to extract image features in Fig. 5 (a). The results of the first block and the features after Pool.3, Pool.4, and Conv.5-3 are merged along the channel direction, and input merged signals into Conv. 6-1. The signals after Conv. 6-2 are transformed using the pyramid pooling module proposed in [40], and the resultant signals are resized to the same size as the signals after Conv. 6-2. Finally, the resized signals and those after Conv. 6-2 are merged along the channel direction, and processed through Conv. 7-1, 2, and 3.

2) Loss Function

The loss function of the proposed CNN method assigns high and medium weights for salient and object regions, respectively, and by contrast, low weights to background regions because they are generally uninteresting. The loss function L is formulated as

$$L_{(\boldsymbol{w})} = \frac{1}{N}\sum_{i=1}^{N}\left\| \frac{\frac{\phi_{(}\boldsymbol{x}_i)}{\max \phi_{(}\boldsymbol{x}_i)} - \boldsymbol{y}_i}{\beta - (\boldsymbol{O}_i + \boldsymbol{y}_i)} \right\|, \tag{5}$$

where $\boldsymbol{y}_i$, $\boldsymbol{x}_i$, $\boldsymbol{O}_i$, $\phi(\cdot)$, and β denote true saliencies, estimated saliencies, object region masks, a normalization function, and a free parameter, respectively. The masks are produced by binarizing signals of the instance segmentation results. $\phi(\cdot)$ normalizes the estimated saliency values in $[0, 1]$. We generally set β to 2 or a value that is the maximum of $\boldsymbol{O}_i+\boldsymbol{y}_i$. If the i-th pixel is in a salient object, $\beta-(\boldsymbol{O}_i + \boldsymbol{y}_i)$ is a low value, and hence this pixel is assigned a high weight.

3) Training

For training, the loss function in Section V-B2 and the training dataset of COCO and SALICON were used [46], [47]. COCO contains natural images and their segmentation masks, and SALICON has saliency maps that correspond to them. The maps were binarized using a threshold value of $\tau = 0.15$, and their elements corresponding to background pixels, which were detected by the masks, were set to zero.

TABLE 8: Scores for the estimation of the importance degree for the PASCAL-S dataset [10].

Image	Spearman's [33]		Kendall [34]		Prop. metric	
	RSOD [26]	Prop.	RSOD [26]	Prop.	RSOD [26]	Prop.
Mini bike	0.869	**0.936**	0.908	**0.956**	0.925	**0.948**
Car	0.146	**1.000**	0.000	**1.000**	0.471	**0.902**
Dog	0.146	**1.000**	0.000	**1.000**	0.261	**1.000**
Horse	0.000	**0.834**	0.000	**0.908**	0.321	**0.822**
Traffic	0.380	**1.000**	0.000	**1.000**	0.402	**0.971**
Average	0.372	**0.419**	0.300	**0.327**	0.457	**0.467**

TABLE 9: Scores for the estimation of the importance degree for the SALICON-based dataset [46], [47].

Image	Spearman's [33]		Kendall [34]		Prop. metric	
	RSOD [26]	Prop.	RSOD [26]	Prop.	RSOD [26]	Prop.
Parking	0.574	**0.607**	0.698	**0.786**	0.461	**0.779**
Party	0.394	**0.555**	0.691	**0.741**	0.430	**0.698**
Woman	0.000	**0.583**	0.000	**0.754**	0.118	**0.501**
Bag	0.028	**0.815**	0.000	**0.887**	0.128	**0.865**
Man	0.586	**0.701**	0.758	**0.887**	0.511	**0.859**
Baseball	0.868	**0.901**	0.900	**0.947**	0.793	**0.924**
Average	0.321	**0.562**	0.434	**0.490**	0.478	**0.507**

Stochastic gradient descent was used as optimizing, where Nesterov momentum, the weight decay, and the learning rate were set to 0.9, 0.5, and 10^{-3}, respectively [51]. β in the loss function was set to 2.3, which was experimentally determined from the ratios of salient, object, and background regions.

C. ESTIMATION OF THE IMPORTANCE DEGREE

In the proposed method, the estimation block process is defined similarly to the proposed procedure in Section IV-D. Object contours are already detected in the first block and their saliency values are estimated in the second block. In the third block, the values within one object contour are summed and the result is its score of the importance degree as given in (4). Similar to the proposed procedure, SOD-ID maps are created based on the resultant scores and linearly quantized with N.

VI. SIMULATION

In this section, we compare the performance of the proposed method and state-of-the-art methods for SOD and SOD-ID. We present the comparisons in Section VI-B and VI-C, respectively, and before that, we discuss the validity of the proposed metric in Section VI-A by presenting some examples. For this simulation, we used the instance segmentation method proposed in [41] in the first block of the proposed method because it is not recent but has high accuracy. Based on Section IV-D, we introduced a dataset from the test sets of COCO and SALICON, which contain images with segmentation masks and their SD maps, respectively, where the proposed dataset is called a SALICON-based dataset in this section. Note that the proposed method is also represented by Prop. in this section.

A. VALIDITY OF THE PROPOSED METRIC

As mentioned in Section IV-C, the validity of the proposed metric is briefly shown in this section. Table 4 shows scores of pairs of arbitrary vectors in Spearman's and Kendall rank correlation coefficients, and the proposed metric. In Table 4, the pairs from the top to the bottom, respectively, indicate various scenarios as follows: same rank and slightly different value, slightly different rank and value, same rank and quite different value, and quite different rank and slightly different value. As mentioned in Section IV-C, SOD-ID metrics have to simultaneously evaluate the rank correlation and the value similarity. In that sense, from the first and third pairs, the proposed metric only satisfies the above property. We observed from the second and fourth pairs that the Kendall coefficient is too sensitive to the rank difference to be used as the SOD-ID metric. The fourth pair shows that the rank correlation is quite different, but its values are almost the same and hence the importance of objects is also considered to be comparable. However, the score obtained using Spearman's coefficient is rather bad and its weight for the rank correlation and the value similarity has been shown to be unbalanced. The proposed metric is clearly more suitable to be used as the SOD-ID metric than the two coefficients.

B. COMPARISON OF THE PROPOSED METHOD WITH SOD METHODS

Settings: HDCT [15], RFCN [16], DHS [21], DSSOD [25], and RSOD [26] were used as SOD methods for comparison. The methods were applied to the test sets of the DUTS, PASCAL-S, and SALICON-based datasets [10], [45]–[47], and the results were evaluated using the F-measure [52]. To calculate the F-measure, it is required that the saliency maps of the PASCAL-S and SALICON-based datasets, and the results of the methods are binarized. Because we set $N = 7$ in this paper and the maps of PASCAL-S have integer values in $[0, 255]$ with $N = 13$, objects whose scores of the importance degree were one or more were recognized as salient for the SALICON-based dataset and therefore the maps of PASCAL-S were binarized with a threshold value of 36. According to the above, the results of HDCT, RFCN, DHS and DSSOD were binarized with a threshold value of 0.14, and for RSOD and Prop., the value was 1.

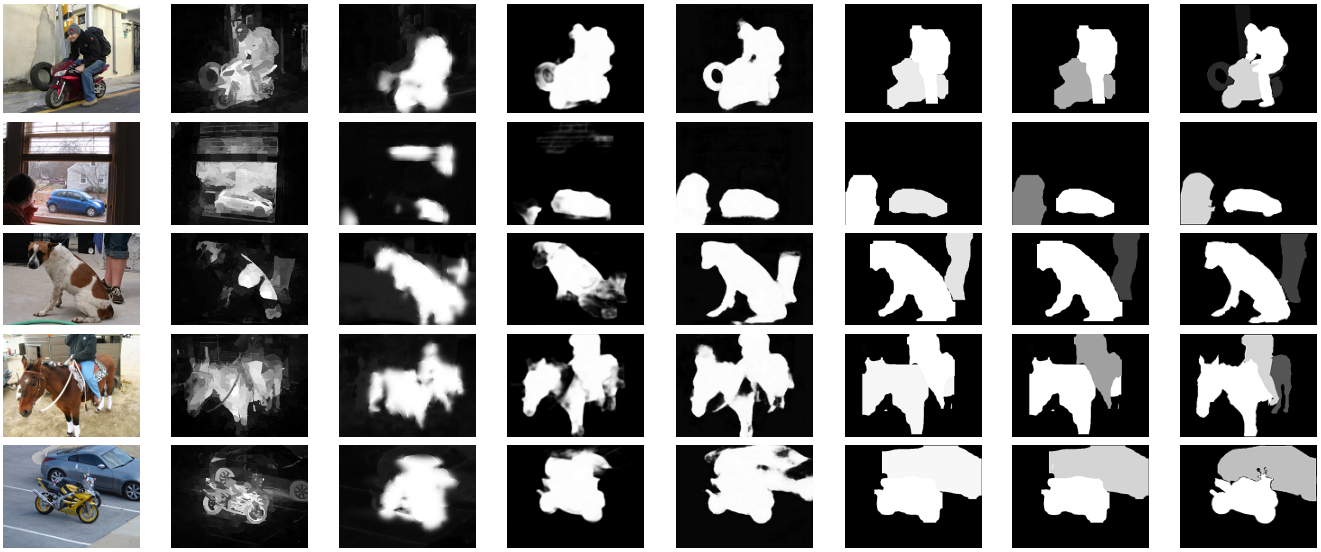


FIGURE 6: Resultant saliency maps for the PASCAL-S dataset [10].

Evaluation: Tables 5, 6, and 7 show the F-measure scores of the methods for each dataset, and Figs. 6 and 7 show the images, their GT maps, and their results before thresholding, where "Average" denotes the average values over all the images in each dataset. The images in the PASCAL-S and SALICON-based datasets generally contain multiple objects, and by contrast, those in the DUTS dataset generally contain one large object or two objects. From Table 5, unfortunately, Prop. had worse scores for DUTS. However, Prop. outperformed the other methods in Tables 6 and 7, and we observed in Figs. 6 and 7 that Prop. accurately estimated object contours. Particularly, Prop. suppressed the inaccurate estimation in "Parking" and "Party" in Fig. 7. Unfortunately, the instance segmentation method often detected nothing for DUTS because of its above characteristic, as shown in the upper half of Table 5. However, the results of Prop. except that case were equivalent to those of the other methods. Prop. can solve this problem using an efficient instance segmentation method that accurately detects objects.

C. COMPARISON OF THE PROPOSED METHOD WITH SOD-ID METHOD

Settings: In SOD-ID, Prop. was compared with RSOD which is the only existing SOD-ID method. The methods were applied to the PASCAL-S and SALICON-based datasets, and the results were evaluated using Spearman's and Kendall rank correlation coefficients and the proposed metric (2), where α and σ were experimentally set to 0.5 and 2.0, respectively. Clearly, the GT and resultant maps were uniformly normalized with $N = 7$.

Evaluation: Tables 8 and 9 show the scores of RSOD and Prop. in the metrics for each dataset, and Figs. 6 and 7 show the images and GT maps in the datasets, and their resultant maps, where high values of pixels in the maps indicate high scores of the importance degree. Note that the rows in Tables 8 and 9 correspond to those in Figs. 6 and 7, respectively. From Table 8 and 9, Prop. clearly outperformed RSOD in terms of the metrics. From Figs. 6 and 7, Prop. accurately estimated the importance degree of objects. Particularly, in "Party," "Woman," and "Man," Prop. estimated the importance degree of small objects that had low saliency scores and were located in highly salient objects.

VII. CONCLUSION

In this paper, we introduced SOD-ID via discussing its definition, significance, dataset condition, and evaluation metric property, and proposed its dataset, metric, and method. The proposed metric consists of the Kendall rank correlation coefficient and mean squared error, and simultaneously evaluates the rank correlation and value similarity for SOD-ID. The proposed dataset is generated using the proposed procedure based on the COCO and SALICON datasets. The proposed method of SOD-ID consists of three processing blocks: instance segmentation, SD, and importance degree estimation. We proposed a CNN-based SD method for the second block that uses the results of the first block. With this strategy, the proposed method objectively outperformed state-of-the-art methods with respect to SOD and achieved an accurate SOD-ID.

ACKNOWLEDGMENT

We thank Irina Entin, M. Eng., and Maxine Garcia, Ph.D., from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript. This work was supported by JSPS KAKENHI Grant Number 16K18104.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Patt. Anal. Mach. Intel., vol. 20,

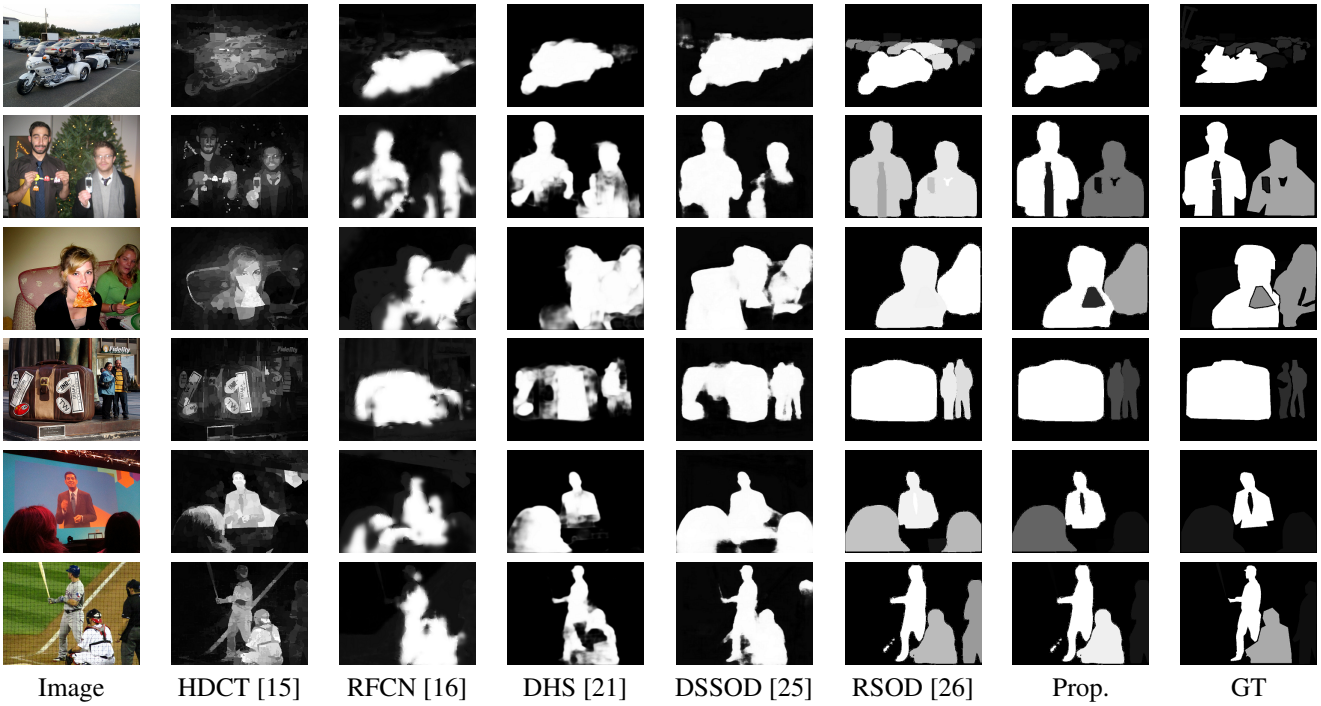


FIGURE 7: Resultant saliency maps for the SALICON-based dataset [46], [47].

no. 11, pp. 1254–1259, 1998.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in Proc. Neural Information Process. Syst., 2006, pp. 545–552.

[3] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2007, pp. 1–8.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2009, pp. 1597–1604.

[5] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: A boolean map approach," IEEE Trans. Patt. Anal. Mach. Intel., vol. 38, no. 5, pp. 889–902, 2016.

[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in Proc. IEEE Int. Conf. Patt. Recognit. IEEE, 2016, pp. 3488–3493.

[7] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," IEEE Trans. Image Process., vol. 26, no. 9, pp. 4446–4456, 2017.

[8] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2012, pp. 733–740.

[9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2013, pp. 2083–2090.

[10] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2014, pp. 280–287.

[11] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," IEEE Trans. Image Process., vol. 24, no. 5, pp. 1639–1649, 2015.

[12] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2015, pp. 1265–1274.

[13] N. Tong, H. Lu, X. Ruan, and M. H. Yang, "Salient object detection via bootstrap learning," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2015, pp. 1884–1892.

[14] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2015, pp. 110–119.

[15] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," IEEE Trans. Image Process., vol. 25, no. 1, pp. 9–23, 2016.

[16] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in Proc. European Conf. Comput. Vis. Springer, 2016, pp. 825–841.

[17] T. Wang, L. Zhang, H. Lu, C. Sun, and J. Qi, "Kernelized subspace ranking for saliency detection," in Proc. European Conf. Comput. Vis., 2016, pp. 450–466.

[18] L. Zhang, C. Yang, H. Lu, X. Ruan, and M.-H. Yang, "Ranking saliency," IEEE Trans. Patt. Anal. Mach. Intel., vol. 39, no. 9, pp. 1892–1904, 2016.

[19] C. Sheth and R. V. Babu, "Object saliency using a background prior," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 1931–1935.

[20] F. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," IEEE Trans. Patt. Anal. Mach. Intel., 2016.

[21] N. Liu and H. J, "Dhsnet: Deep hierarchical saliency network for salient object detection," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2016, pp. 678–686.

[22] L. Zhang, J. Ai, B. Jiang, H. Lu, and X. Li, "Saliency detection via absorbing markov chain with learnt transition probability," IEEE Trans. Image Process., vol. 27, no. 2, pp. 987–998, 2017.

[23] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2017, pp. 2386–2395.

[24] J. Yang and M. H. Yang, "Top-down visual saliency via joint crf and dictionary learning," IEEE Trans. Patt. Anal. Mach. Intel., vol. 39, no. 3, pp. 576–588, 2017.

[25] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2017, pp. 3203–3212.

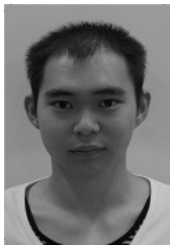
[26] M. Amirul Islam, M. Kalash, and N. D. B. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2018, pp. 7142–7150.

[27] C. Aytekin, A. Iosifidis, and M. Gabbouj, "Probabilistic saliency estimation," Patt. Recognit., vol. 74, pp. 359–372, 2018.

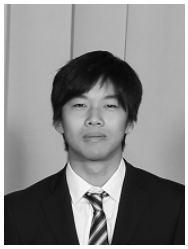
[28] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2019, pp. 6103–6112.

[29] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual

saliency detection with applications to image thumbnailing," in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 2232–2239.
[30] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," ACM Trans. graphics, vol. 29, no. 6, pp. 160–9, 2010.
[31] A. Mansfield, P. Gehler, L. V. Gool, and C. Rother, "Scene carving: Scene consistent image retargeting," in Proc. European Conf. Comput. Vis. Springer, 2010, pp. 143–156.
[32] A. Jose and I. Heisterklaus, "Bag of fisher vectors representation of images by saliency-based spatial partitioning," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2017, pp. 1762–1766.
[33] C. Spearman, "The proof and measurement of association between two things." 1961.
[34] M. G. Kendall, "A new measure of rank correlation," Biometrika, vol. 30, no. 1/2, pp. 81–93, 1938.
[35] N. İmamoğlu, C. Zhang, W. Shmoda, Y. Fang, and B. Shi, "Saliency detection by forward and backward cues in deep-cnn," in Proc. IEEE Int. Conf. on Image Process., 2017, pp. 430–434.
[36] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "Salnet360: Saliency maps for omni-directional images with cnn," Elsevier Trans. Signal Process. Image Communication, vol. 69, pp. 26–34, 2018.
[37] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Inner and inter label propagation: Salient object detection in the wild," IEEE Trans. Image Process., vol. 24, no. 10, pp. 3176–3186, 2015.
[38] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 991–998.
[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2015, pp. 3431–3440.
[40] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2017, pp. 2881–2890.
[41] B. Romera-Paredes and P. H. S. Torr, "Recurrent instance segmentation," in Proc. European Conf. Comput. Vis. Springer, 2016, pp. 312–329.
[42] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2017, pp. 2359–2367.
[43] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2014, pp. 891–898.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations, 2015.
[45] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2017, pp. 136–145.
[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Proc. European Conf. Comput. Vis. Springer, 2014, pp. 740–755.
[47] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2015, pp. 1072–1080.
[48] K. Li, B. Hariharan, and J. Malik, "Iterative instance segmentation," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2016, pp. 3659–3667.
[49] A. Arnab and P. H. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in Proc. IEEE Conf. Comput. Vis. Patt. Recognit., 2017, pp. 441–450.
[50] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proc. International Conf. on Mach. learn., 2010, pp. 807–814.
[51] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[52] R. Szeliski, Computer Vision: Algorithms and Applications. Springer London, 2010.



YO UMEKI received the B. Eng. and M. Eng. degrees from Nagaoka University of Technology, Nagaoka, Japan, in 2015 and 2019, respectively. He is currently a doctoral student with the Department of Information Science and Control Engineering. His main research interest is saliency detection.



ISANA FUNAHASHI Isana Funahashi received the B. Eng. and M. Eng. degrees from Nagaoka University of Technology, Nagaoka, Japan, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer and Network Engineering with the University of Electro-Communications, Tokyo, Japan. His research interests are image processing and computer vision.

TAICHI YOSHIDA (M'15) received the B. Eng., M. Eng., and Ph.D. degrees in Engineering from Keio University, Yokohama, Japan, in 2006, 2008, and 2013, respectively. In 2014, he joined Nagaoka University of Technology. In 2018, he joined the University of Electro-Communications where he is currently an Assistant Professor in the Department of Communication Engineering and Informatics. His research interests are filter bank design and image coding applications.

MASAHIRO IWAHASHI (SM'08) received the B. Eng., M. Eng., and D. Eng. degrees in Electrical Engineering from Tokyo Metropolitan University in 1988, 1990, and 1996, respectively. In 1990, he joined Nippon Steel Co., Ltd. From 1991 to 1992, he was seconded to Graphics Communication Technology Co., Ltd. In 1993, he joined Nagaoka University of Technology, where he is currently a Professor in the Department of Electrical Engineering, Faculty of Technology. From 1995 to 2001, he was also a lecturer at Nagaoka Technical College. From 1998 to 2001, he relocated to Thammasat University, Thailand, and to the Electronic Engineering Polytechnic Institute of Surabaya, Indonesia, as a JICA expert. His research interests are digital signal processing, multi-rate systems, and image compression. He served as an editorial committee member of the IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences from 2007 to 2011.