IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2022, accepted 13 August 2022, date of publication 19 August 2022, date of current version 29 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3200486

RESEARCH ARTICLE

Saliency-Based Multiple Region of Interest Detection From a Single 360° Image

YUUKI SAWABE[1], (Graduate Student Member, IEEE), SATOSHI IKEHATA[2], (Member, IEEE), AND KIYOHARU AIZAWA[1], (Fellow, IEEE)

[1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8654, Japan
[2]National Institute of Informatics, Tokyo 101-8430, Japan

Corresponding author: Yuuki Sawabe (sawabe@hal.t.u-tokyo.ac.jp)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT 360° images are informative – it contains omnidirectional visual information around the camera. However, the areas that cover a 360° image is much larger than the human's field of view, therefore important information in different view directions is easily overlooked. To tackle this issue, we propose a method for predicting the optimal set of Region of Interest (RoI) from a single 360° image using the visual saliency as a clue. To deal with the scarce, strongly biased training data of existing single 360° image saliency prediction dataset, we also propose a data augmentation method based on the spherical random data rotation. From the predicted saliency map and redundant candidate regions, we obtain the optimal set of RoIs considering both the saliency within a region and the Interaction-Over-Union (IoU) between regions. We conduct the subjective evaluation to show that the proposed method can select regions that properly summarize the input 360° image.

INDEX TERMS 360° image, saliency, region of interest, virtual reality technology, data augmentation.

I. INTRODUCTION

In recent years, 360° images and videos have attracted significant attention owing to their advantage in terms of their omnidirectional information over the perspective images whose typical field of view (FoV) is less than 65° (*i.e.*, normal field-of-view (NFoV) [1]). However, the human vision system does not have such a large FoV and achieving an efficient way of viewing the 360° image is an important problem in preventing important information in an 360° image from being overlooked.

To tackle this problem, one effective solution is to summarize the entire 360° image into a set of Region of Interest (RoI) and present them to the viewer when browsing 360° images or videos [1], [2]. In our context, RoI is defined as the local region of a small FoV that might draw the attention of the viewer, and is expected to represent the important information in the input 360° image or video. In the early work by Su *et al.* [1], which first attempted to extract an RoI from a 360° video, it was assumed there was only a single RoI in each frame, which can be problematic when multiple important items are observed at the same time. In addition, the input must have multiple frames from a 360° video because it utilizes temporal information to decide which view is good to crop. To tackle this issue, Xiong *et al.* [2] have recently proposed a method for extracting multiple RoIs from a single 360° image by projecting the 360° image onto a unit cube and optimizing the rotation of the cube such that the extracted RoIs are placed at the centers of its faces. Although it allows multiple RoIs from a single image, the FoV covered by each face of the cube (*i.e.*, 90°) is larger than human's FoV and the number (*i.e.*, six) and relative positions of the RoIs (*i.e.*, two different faces of cubes) are always fixed. Therefore, it is highly likely that the extracted RoIs will be either of redundant or scarce.

To properly support a more effective 360° image browsing, we propose a method that takes a single 360° image as input and outputs the fixed number of RoIs *whose positions, and*

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

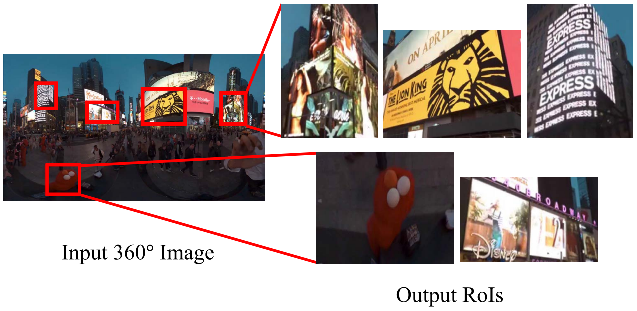


FIGURE 1. The task proposed in this study. Detecting the Region of Interest on the right from the 360° ERP image on the left.

corresponding FoVs are adaptively selected. To the best of our knowledge, this is the first attempt to predict multiple NFoV RoIs from a single 360-degree image without constraining their sizes and positions.

Example results are illustrated in Fig. 1. In this example, we extracted five RoIs of varying size from the input 360° image in ERP (EquiRectangular Projection) format. Our method firstly divides the input ERP image into multiple overlapping candidate rectangular regions through Selective Search [3] and find the optimal fixed-sized (*i.e.*, five in this example) subset by optimizing our new evaluation function, namely, *Salient-IoU*, which considers the saliency values within each region and the IoU which measures the overlap between two regions. Saliency [4] is a quantitative measure of attracting human visual attention and has been used in image recognition, object detection, robotics, and advertising design. In our framework, the accurate prediction of saliency values is a critical component. The prediction of saliency originated from Itti *et al.*'s method [5], which was the first to introduce a bottom-up model of human visual attention, and in recent years, deep neural network models based on gaze information have become mainstream [6]. However, existing saliency map (a 2-D image that contains the saliency value at each pixel) prediction is mainly for perspective projection cameras, and the performance for 360° images is fairly limited [7], [8], [9], [10]. To overcome this limitation, we propose a simple but effective new data augmentation method using a random rotation on a unit sphere which deals with the scarce, strongly biased training data of existing single 360° image saliency prediction dataset [11], [12].

Owing to a lack of existing research on the multiple RoI extraction from a single 360° image for the summarization purpose, we evaluate our method based on the user study on Amazon Mechanical Turk [13] and show that our results match the human's intuition.

A summary of our contributions is as follows:

- To the best of our knowledge, this is the first work that predicts multiple RoIs from a single 360-degree image without constraining the size and position of the RoIs.
- We present the simple but effective data augmentation strategy for improving single 360° image saliency prediction by which our saliency prediction significantly outperforms the state of the art.
- We conducted a user evaluation on the Amazon Mechanical Turk to illustrate the performance of our RoI prediction.

II. RELATED WORKS

A. 360° INFORMATION SUMMARIZATION

It is an important task to efficiently present 360° images/ videos to the limited human's view. However, as already mentioned, there is no prior work which predicted multiple RoIs from a single 360° image, to the best of knowledge. On the other hand, the summarization of 360° video had been an active topic that is to present the most important view direction at each frame. For instance, Su *et al.* [1] proposed an optimization-based algorithm to find a path over the spatio-temporal glimpses that maximize the accumulated capture-worthiness score while obeying a smooth camera motion constraint. This work was later extended to allow more general camera control such as zooming [14]. Benefiting from deep-learning-based object detection methods, Deep 360 Pilot [15] presented an object-centric deep-learning-based agent for piloting through 360° sports videos automatically. While typical 360° video summarizing task targets to find the optimal spatial camera trajectories, Lee *et al.* [16] also addressed story-based temporal summarization by leveraging the memory networks.

Recently, Wang *et al.* [17] have presented *Transitioning360*, a tool for 360° video navigation on 2-D displays by transitioning between multiple NFoV views that track potentially interesting targets or events. They combined saliency map, optical flow and object instances computed from the input 360° video and optimized the virtual NFoV paths based on both contents and temporal smoothness. While this work also uses the saliency information to detect important regions in 360° video and presented viewers multiple RoIs at the same time, the predicted RoIs are basically object centric, which relies on the specific object categories and more importantly this method cannot be applied to a single frame. On the other hand, our method introduces Selective Search [3] to find perceptually important, non-object-centric candidate regions and is completely applicable to a single 360° image.

B. SALIENCY MAP PREDICTION FROM 360° IMAGE

In contrast to the saliency map prediction on normal perspective images, estimating the saliency map of a 360° image in ERP format is more challenging due to distortions caused by projection from a sphere to a plane.

Traditionally, the existing bottom-up model of human visual attention [5] was extended to 360° images in ERP format and the saliency map was predicted based on that model [18]. However the method that only considers low level visual features limits its prediction accuracy.

The data-driven, image-based saliency detection was made possible by the advent of public data sets. The first moderate-scale public datasets of 360° images with

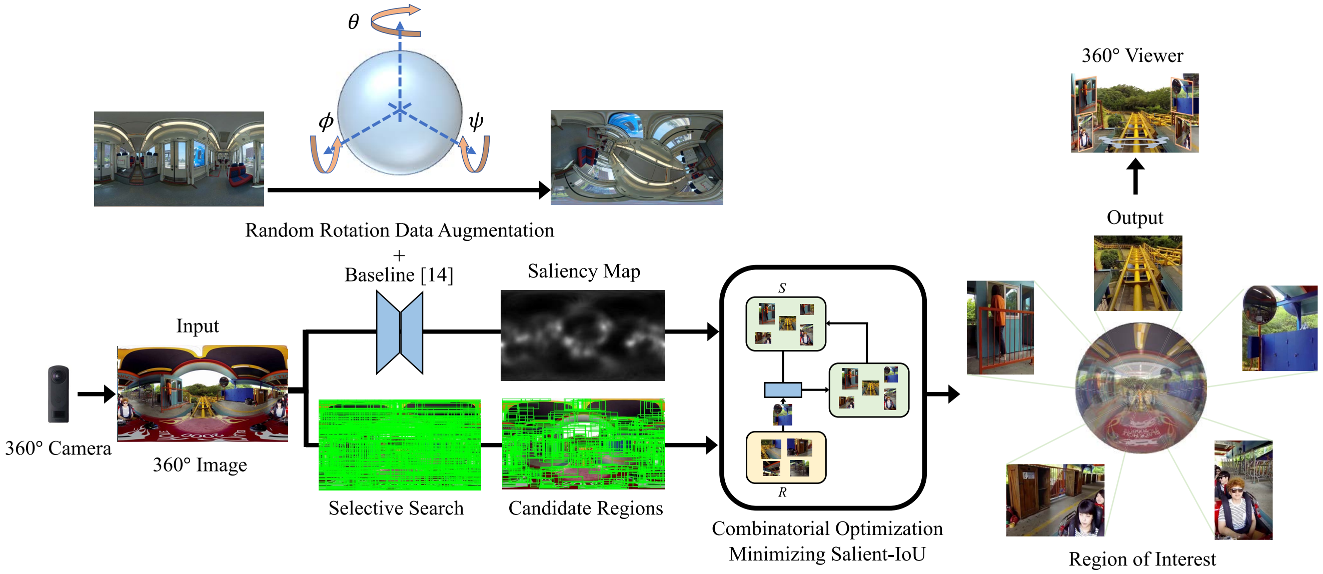


FIGURE 2. A framework for detecting multiple Region of Interest from a single360° image.

associated eye and head movement data had been presented by Rai *et al.* [19] which consists of sixty different 360° images and gaze information by at least 40 observers. This dataset was later followed up by Erwan *et al.* [20] to extend it to 360° videos. Using these dataset, Monroy *et al.* [9] presented the first data-driven saliency map prediction method which takes a 360° image as input and splits it into six patches to be fed to convolutional neural networks (CNN). Chen *et al.* [21] presented a spatio-temporal nework to predict 360° video saliency with cube-padding technique to avoid the sphere-to-plane distortion problem. Zhang *et al.* [22] presented the 360° video saliency detection by a spherical convolution neural network trained on 104 360° videos viewed by 27 human subjects. Chao *et al.* [7] extended the perspective saliency prediction model trained with adversarial examples [23] to 360° images and won Salient360! Grand Challenges at ICME'18 in the task of prediction of head and eye saliency, and this method became a touchstone in the domain of the 360° image saliency map prediction.

There are two most recent works leveraging state-of-the-art deep learning techniques. Haoran *et al.* [10] presented the 360° saliency detection algorithm based on the graph convolutional neural networks. Specifically, a spherical graph signal was constructed from a ERP image and saliency map was generated from the spherical features on the graph. Martin *et al.* [8] leveraged 360°-aware convolutions that represent kernels as patches tangent to the sphere where the panorama is projected, and a spherical loss function that penalizes prediction errors for each pixel depending on its coordinates in a gnomonic projection.

One of the remaining challenges is the limited amount of data compared to perspective images. In particular, data diversity is an important issue, and is also a cause of strong center bias, where the salient region is concentrated on the equator. However, no clear solution to this problem has been proposed yet, even with the state-of-the-art methods described above [7], [8]. In this work, we show through experiments that this can be addressed by the data augmentation by rotating the ERP image in spherical coordinates added on top of the existing architecture [8].

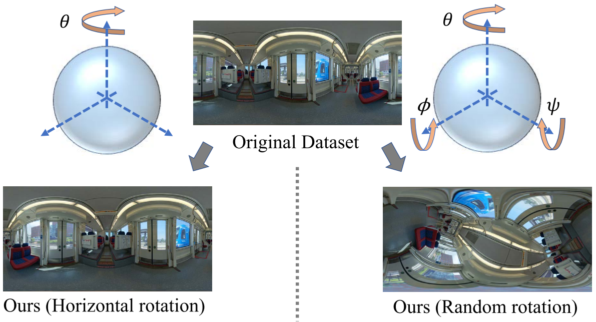


FIGURE 3. Examples of random rotation data augmentation.

III. PROPOSED METHOD

An overview of the proposed method is illustrated in Fig. 2. Given a single 360° image in ERP as input, the proposed method (1) predicts a saliency map using prior baseline networks [8] trained using our data augmentation technique by spherical random rotations, (2) extracts RoI candidates based on Selective Search [3], and (3) optimizes a set of RoIs based on our Salient-IoU evaluation score.

A. SALIENCY MAP PREDICTION USING SPHERICAL RANDOM ROTATION AUGMENTATION

Based on its definition, the RoIs draw more attention from human viewers than other regions in the same image – RoIs can be considered as the region of local maximum saliency. As introduced in Section II-B, there are a few deep neural network models for this task but trained on mid-scale datasets [12], [19] which contain less than two hundred pairs of a 360° image and a corresponding gaze information where most salient regions are concentrated near the equator. Näively training a network on this training data inevitably results in this strong center bias and makes it difficult to extract RoIs apart from the equator.

To tackle both the problems of scarce of data and the strong center bias, we propose a simple but effective spherical data augmentation strategy. As illustrated in Fig. 3, given a pair of 360° images and gaze maps in ERP format, we first back-project ERP images onto the unit sphere, then apply random rotation, and project them back onto the ERP coordinates. In this random rotation, the image with its paired saliency map on sphere is randomly rotated around three axes individually: $\theta \in [-\pi, \pi]$ around the gravity direction, $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ around the grazing axis, and $\psi \in [-\pi, \pi]$ around the axis perpendicular to both rotation axes. After the rotation, the salient region around the equator moves away from the equator; therefore, the strong center bias is removed. In our experiments, we will show that simply applying the proposed method to the state-of-the-art single image 360° saliency prediction model by Martin *et al.* [8] dramatically improves the prediction accuracy even with mid-scale training examples.

B. RoI EXTRACTION THROUGH SALIENT-IoU OPTIMIZATION

The goal of our framework is to extract a predefined number of RoIs (*e.g.*, $n = 5$ in this study) as distinctively salient regions in the input 360° image. However, most pixels in the predicted saliency map contain non-smooth, non-zero entries and simply thresholding the saliency values will result in the generation of a number of isolated regions. Another possible strategy can be to first apply an object detection algorithm (*e.g.*, [24]) to the input image and simply take top-n objects with the highest saliency values (*i.e.*, salient object detection). Nevertheless, this simple strategy cannot extract an RoI that includes multiple objects, and more importantly, the result is restricted to specific recognizable categories of objects. Instead, we take a two-step approach, which is composed of the extraction of candidate regions that are perceptually important and the optimization of selection of subset of RoIs that satisfy our criteria.

In the first step, we apply Selective Search [3] algorithm to the input 360° image to extract the candidate regions. Selective Search, which is used for producing object proposals in the early time of object detection algorithms [24], greedily merges superpixels based on low-level features such as the color, texture, size, and fitness to extract perceptually coherent regions. Selective Search is a purely bottom-up approach that works for various scenes, unlike recent top-down region proposal neural networks [25] which prefer regions around specific recognizable categories of objects. The output of the Selective Search is illustrated in Fig. 4. The resultant region proposals are overlapping rectangles of various sizes distributed over an entire image. Because we assume that RoIs whose corresponding FoV is smaller than NFoV, we exclude regions where the latitude or longitude FoV is larger than NFoV (*i.e.*, 65°) from the result of Selective Search.

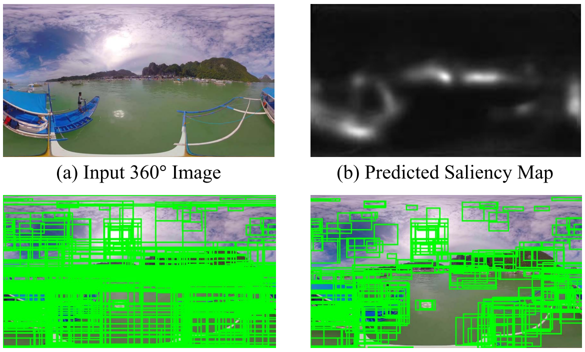

(a) Input 360° Image
(b) Predicted Saliency Map
(c) Candidate Regions
(d) Regions thresholded with NFoV

FIGURE 4. Top row: Inputs for the RoI extraction. Bottom row:Before and after thresholding output regions of Selective Search in NFoV=65°.

Using the predicted saliency map, our task is then to find the optimal subset of the candidate regions that maximally summarizes the entire 360° image in a perceptually plausible manner. This is a non-trivial problem because if we only use top-n regions of the average or summation of the saliency values, a number of overlapping regions will be extracted around the most salient area in the entire image. It is not necessarily desirable to extract regions only around the most salient area because the other important areas are most likely to be overlooked. Assume we want to extract n RoI regions, it is desirable that they are different top-n salient regions from different parts of the image.

Then, we propose to extract of multiple saliency regions with a certain size that have less overlap with each other. Specifically, we minimize the cost function, namely Salient-IoU (γ) which evaluates a subset of n regions (S) from a set of all candidate regions (R) generated by Selective Search. SIoU considers both the sum of the saliency of the region and the overlap between the regions and gives a smaller value to a subset that is considered to be better as follows:

$$\gamma(S) = \frac{a}{n}\sum_{i=1}^{n}\frac{1}{g(I_i)} + (1-a)\frac{1}{{}_nC_2}\sum_{i,j\in[1,n],i\neq j}\mathrm{IoU}(I_i, I_j), \quad (1)$$

where $I_{k\in[1,n]} \in S$ is the region included in S, and $g(I_k)$ is the sum of the predicted saliency values in the region. Owing to the projection distortion of the ERP format, the local summation of saliency values is overestimated in high-latitude regions. Therefore, we multiply $w = \cos\lambda$ using the predicted saliency map, where λ is the latitude $\lambda \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ used to maintain a constant pixel density on a surface of a unit sphere as presented in [26]. The weighted saliency map is then ℓ_2-normalized such that the summation of all saliency values become one. $IoU(I_i, I_j)$ is the Intersection-Over-Union

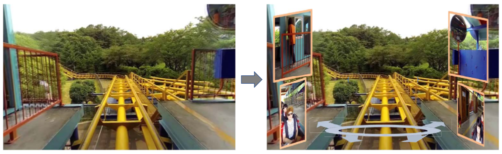


FIGURE 5. An example of 360° image viewer with the RoIs.

(IoU) [24], which becomes one when two regions are completely overlapped and zero when they are not, and a is the balancing weight ($a \in [0, 1]$) which controls the contribution of the saliency and the IoU. When a is close to 1, regions of higher saliency are preferred, and when a is close to zero, overlapping regions are less.

The algorithm used to find the optimal set of RoIs based on Salient-IoU minimization is as follows: Given a collection of regions (R) from Selective Search which was filtered out by their FoV, the region subset S is initialized with the top n candidate regions that have the highest total saliency values. and the elements in S are removed from R. We then greedily replace one region in S with another region in R, which has the highest total saliency values in R, one by one, and compare values of SIoU before and after the replacement. If the value of SIoU gets smaller after the replacement, S and R are updated – the replacement is accepted and the element is removed from R. In the implementation, S is an array which consists of n regions. With the element of index from zero to $n - 1$ in S, replacement, calculation of SIoU values, and comparison are executed with the current target element which has the highest total saliency values in R. If the all possible replacements between each element in S and the current target element in R are rejected, the current target element in R is removed from R. Next, the region with the highest total saliency value in R becomes the current target. This operation is repeated until there are no more candidates in R, and the final S is the optimal RoI set.

C. 360° IMAGE VIEWER WITH EXTRACTED RoIs

The extracted RoIs can be directly overlaid to the input 360° image, however there should be more effective way to display summarized information to a viewer. For example, by applying the proposed method to each frame of a 360° video, we can efficiently extract RoIs that changes in time series to the observer. While this is out of the main scope of this paper, we designed a mock-up GUI and displayed RoIs extracted from the video frames to the observer as shown in Fig. 5. We empirically confirmed that the proposed GUI can efficiently teach viewers what important items and events exist outside viewer's field of view. The better GUI design based on our RoI extraction method is left for our future work.

IV. EXPERIMENTAL RESULTS

We conducted two main experiments to demonstrate the effectiveness of our proposed method. First, we evaluated our spherical random rotation augmentation for better training of the baseline saliency prediction network. Second, we conducted a user study to evaluate the cognitive appropriateness of obtained RoIs.

A. QUANTITATIVE EVALUATION OF SALIENCY MAP PREDICTION

Dataset details: We used Salient360! [11], [12] dataset for the saliency prediction network training and the evaluation. It contains 85 ERP images and corresponding ground truth saliency maps obtained using the eye tracker. We split entire pairs of 360° image and saliency map into 78 for training and seven (*i.e.*, P91, P93-P98) for the test.

Evaluation metrics: We used six evaluation metrics listed below – five of them were used in the Grand Challenge of ICME'17 and ICME'18 of Salient360! [11] and AUC_Borji [27] was included as well.

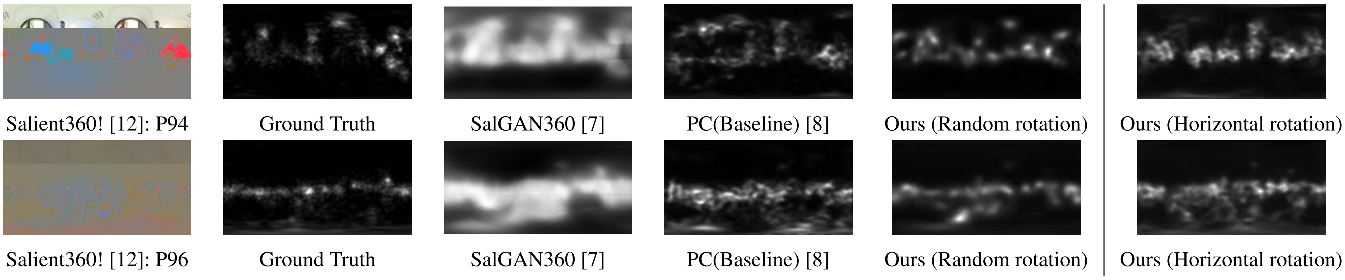
- Normalized Scanpath Saliency (NSS)
- Pearson's Correlation Coefficient (CC)
- Similarity or histogram intersection (SIM)
- Kullback-Leibler divergence (KLD)
- Area under ROC Curve by Judd (AUC_Judd)
- Area under ROC Curve by Borji (AUC_Borji)

Note that we followed the evaluation framework in salient360! [11]. Specifically, predicted saliency maps were weighted by their latitude to avoid overestimating the error in the high-latitude region because the sampling point at high-latitudes on the unit sphere are stretched horizontally on the ERP image.

Implementation Details: We evaluated our data augmentation method by the random spherical rotation using one of the state-of-the-art 360° saliency prediction network by Martin *et al.* (Baseline, PanoramicConv, PC) [8] as the backbone network without changing their original implementation. In addition to our data augmentation technique, Martin *et al.* used seven augmentation methods (*i.e.*, three flips and additive Gaussian noise, Poisson noise, salt-pepper noise and speckle noise). It is a current state of the art, and we used them in the evaluation as well.

These seven augmentations cannot alleviate the problem of the strong center bias of 360° images., which is a special for saliency prediction of 360° images. To remove the effects of the strong center bias, we used our random rotation augmentation in addition to the seven augmentations. During training, the input 360° image in ERP format was downsampled to 256 × 128 and the parameters were optimized with the Momentum SGD [28] and Spherical Mean Squared Error (MSE) Loss [29]. The hyperparameters for the training are as follows: epoch=10000, batch size=32, learning rate=10^{-4}, momentum=0.9, and weight decay=10^{-5}. The entire evaluation framework was implemented using PyTorch [30] where the network was trained and tested on a single NVIDIA Quadro RTX 8000 with 48GB of memory.

Results: The results are shown in Table 1. The prediction by Martin *et al.* without our spherical rotation augmentation is shown as the baseline. Our method was implemented by using it as a backbone and additionally applied random spherical rotation augmentation (Proposal w/ random). For the ablation



Salient360! [12]: P94 | Ground Truth | SalGAN360 [7] | PC(Baseline) [8] | Ours (Random rotation) | Ours (Horizontal rotation)

Salient360! [12]: P96 | Ground Truth | SalGAN360 [7] | PC(Baseline) [8] | Ours (Random rotation) | Ours (Horizontal rotation)

FIGURE 6. From left to right: input image of test data, ground truth of saliency map, results of training with SalGAN360 [7], results of training with the baseline [8], results of training with data augmentation by random rotation of the proposed method, and results of training with data augmentation by horizontal rotation of the proposed method.

TABLE 1. Evaluation of saliency prediction with data augmentation.

	AUC_Judd↑	AUC_Borji↑	NSS↑	CC↑	SIM↑	KLD↓
SalGAN360 [7]	0.758	0.703	**1.309**	0.234	0.466	1.890
PC(Baseline) [8]	0.728	0.693	0.874	0.432	0.545	1.074
Ours w/ random rotation	**0.774**	**0.735**	1.183	**0.584**	**0.609**	**0.842**
Ours w/ horizontal rotation	0.772	0.730	1.105	0.553	0.592	0.949

study, we also showed the result in which the training data was rotated only around the gravity axis (Proposal w/ horizontal). Note that the rotation was only applied to the training data, but not to the test data. Because Martin *et al.* 's and Chao *et al.* [7]'s models' codes are publicly available among the the state of the art methods, we used Chao *et al.* 's model for the comparison with the same setting of Martin *et al.*'s. Papers [10], [31] also used Chao *et al.* 's model as a comparison method.

For all evaluation metrics, the random rotation augmentation was shown to most effectively improve the performance, and its accuracy is significantly better than that of the baseline. We also see that the horizontal rotation is not enough.

Figure 6 shows a visualization of the obtained saliency. It can be confirmed that the baseline method tends to predict high saliency values near the equator owing to the strong center bias of the training data, whereas the proposed method can predict the saliency robustly even within a high latitude region. Besides, even if our horizontal rotation is applied to the dataset, predicted saliency values tend to be blurred and spread within the middle latitude region because of the strong center bias. In contrast, when we use the random rotation, the blurred area becomes denser, and it shows the random rotation is necessary for a better prediction.

B. EVALUATION OF SINGLE 360° IMAGE MULTIPLE RoI PREDICTION

1) DATASET CONSTRUCTION

There is no existing dataset for evaluating multiple RoI prediction task from a single 360° image, therefore we constructed the evaluation dataset which consists of pairs of 360° image and RoIs annotated by human. Some examples in our dataset are shown in Fig. 7.

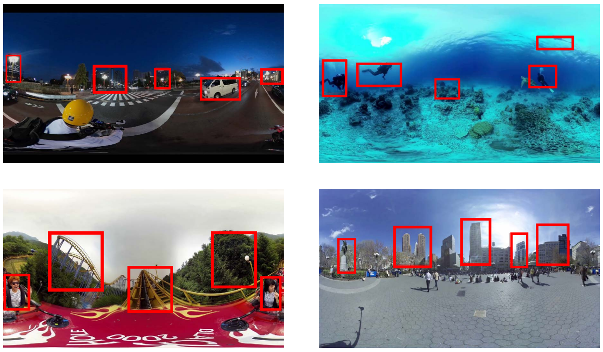


FIGURE 7. Examples of user-annotated RoIs dataset for evaluation.

Our dataset consists of forty five 360° images extracted from five 360° outdoor video clips on YouTube [32] and corresponding RoI annotation by three different persons. To annotate RoIs on each 360° image, we recruited crowd workers using Amazon Mechanical Turk [13] to ask to annotate five impressive regions in each 360° image of the dataset. The dataset construction process is detailed as follow.

First, a cloud worker was instructed to look around an omnidirectional image on the web browser through cropped perspective view by freely changing the view direction and remember the scene. Actual instructions are as follows: (1) Look around the image by dragging the cursor. (2) Remember the parts of the image that are particularly impressive. (3) Take at least 30 seconds. Make sure not missing the ceiling or floor.

Next, we showed the same 360° image but in ERP format to the worker and asked him/her to draw five bounding boxes around impressive parts. The instructions were as follows: (1) Use the bounding box tool to draw boxes around the area that are impressive to you. Draw a rectangle using your mouse

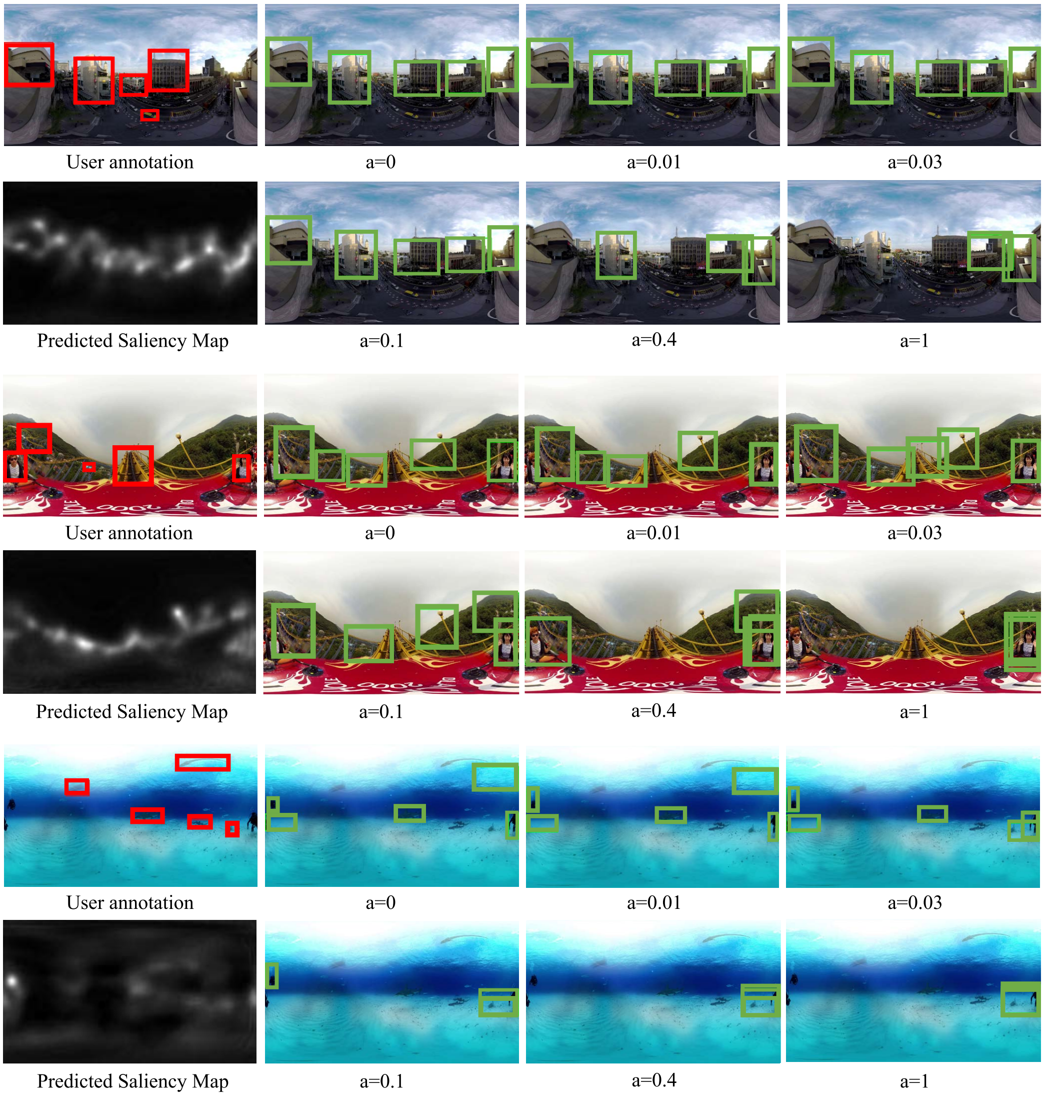


FIGURE 8. **Examples of RoIs obtained by our framework with different *a*s.**

over the area. (2) Each rectangle should be less than 1/4 of the total image. (3) This is not an object detection annotation, so you can enclose areas where there are no objects. It is okay if the areas overlap each other.

A total of 14 crowd workers participated in the annotation. We did not impose any restrictions on the qualifications of the crowd workers, but only approved those workers who spent more than 60 seconds on the task. Our dataset includes 50 images in total, and each image was annotated by 3 persons, resulting in 150 pairs of 360° image and user-annotated RoIs. Excluding the inappropriate 5 images such as a blackout scene, we finally constructed an evaluation dataset of 45 images and 135 RoI annotations. Since RoIs in an image is highly subjective in nature, we didn't merge three annotations of the same 360° image in evaluation.

2) QUALITATIVE EVALUATION

Some examples of our result applied to the constructed dataset are shown in Fig. 8. The predicted RoIs (green color boxes) are overlaid on the input 360° image with different choices of a in Eq. (1). We observed that the prediction results were reasonably close to one of the user-annotated RoIs. As expected, only considering the saliency information (*i.e.*, $a = 1$) gave significantly overlapped RoIs, and we couldn't detect multiple RoIs. On the other hand, spatially distributed RoIs were extracted when we consider both visual saliency and IoU. In failure cases, as shown in the bottom of Figure 9, images with few objects tended to produce RoIs that were significantly different from the user annotation. This is probably due to the fact that the training dataset for the saliency prediction does not include very small salient items.

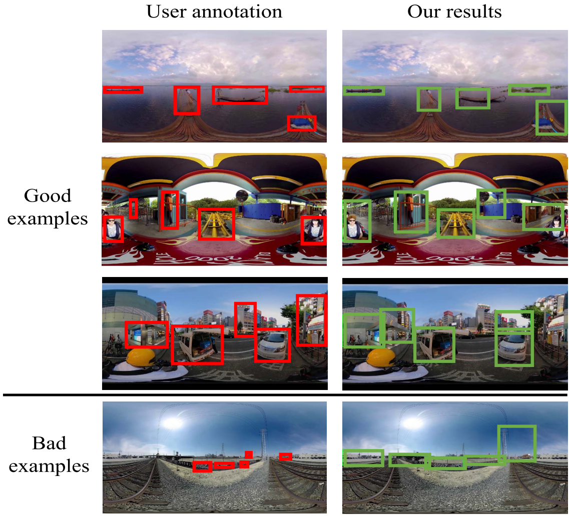


FIGURE 9. Comparison of user annotated RoIs and our detected RoIs.

3) QUANTITATIVE EVALUATION

We conducted a quantitative comparison between user-annotated RoIs (S_{User}) and the RoIs predicted by the proposed method (S_{Pred}). The evaluation metrics for the quantitative comparison were the normalized Euclidean distance (L2 norm) on the image between the center points of the predicted and user-annotated RoIs, and the IoU between them. Considering that the right and left edges of the ERP image are connected, the longest distance is $D = \sqrt{\frac{W}{2}^2 + H^2}$ when the number of vertical pixels in the ERP image is H and the number of horizontal pixels is W. The normalized Euclidean distance D between the regions $\alpha = (x_\alpha, y_\alpha, w_\alpha, h_\alpha)$ and the region $\beta = (x_\beta, y_\beta, w_\beta, h_\beta)$ is computed as follows:

$$D = \frac{\sqrt{min(|x_\alpha - x_\beta|, W - |x_\alpha - x_\beta|)^2 + (y_\alpha - y_\beta)^2}}{\sqrt{\frac{W}{2}^2 + H^2}}. \quad (2)$$

For the IoU, because Selective Search [3] is originally an algorithm for perspective images, it does not take into account a candidate region that crosses the right to the left edge of an ERP image.

We considered the following two methods for paring human-annotated regions and predicted regions for comparison. S_{User} and S_{Pred} means a set of five RoIs by human-annotation and our prediction, respectively.

- For each RoI in S_{Pred}, choose the region of S_{User} that gives the best evaluation, and average all of the best evaluations. (Eval1)
- For each RoI in S_{User}, choose the region of S_{Pred} that gives the best evaluation, and average all of the best evaluations. (Eval2)

Eval1 and Eval2 are similar to precision and recall, respectively. The graphs of the results for these evaluations are shown in Fig. 10. In addition to the human-annotation and our prediction, we show a random selection in Fig. 10. The random selection randomly chose five regions from the regions obtained by Selective Search. Its evaluation is shown by dashed lines in the figures.

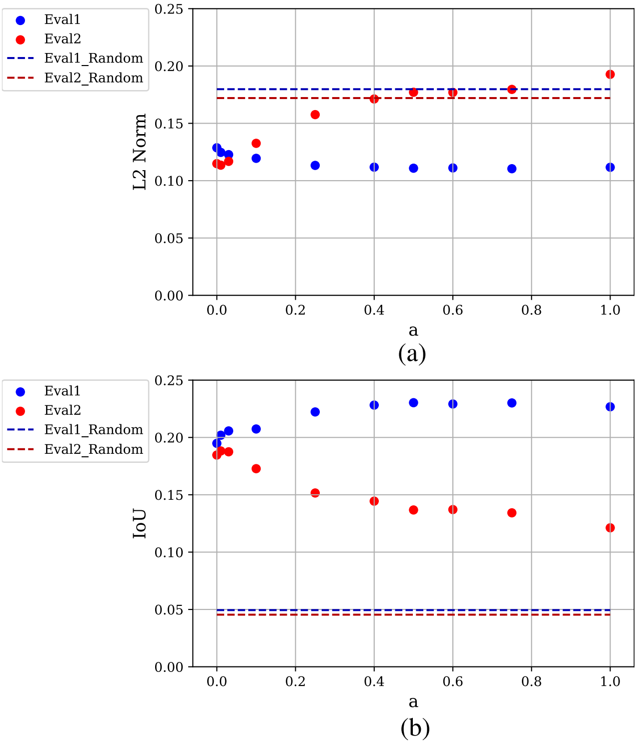


FIGURE 10. Comprisons between the the user annotated RoIs and our detected RoIs for varying *a*s. (a) L2 norm, (b) IoU.

It is noteworthy that when the saliency was more weighted in Salient-IoU, L2 norm slightly decreased in Eval1 and increased in Eval2. This indicates that the RoIs predicted by emphasizing only saliency and neglecting the degree of overlap between RoIs were concentrated in specific high salient region, and the RoIs did not well distribute within the images. In other words, although the prediction with $a = 1$ accurately select the highest salient region, but "overlook" is not avoided. These results are consistent with the observations in the qualitative evaluation presented in the previous section.

C. USER STUDY FOR EVALUATING PREDICTED RoIs

We conducted subjective evaluation to verify the quality of predicted RoIs. For presenting predicted/annotated RoIs on the NFoV perspective image, we converted RoIs defined on the ERP coordinates to ones on perspective images without projection distortions. In this conversion, we set a tangent plane contacting on the unit sphere at the center coordinate of the RoI on the ERP image, and project points on the unit sphere using the RoI as the projection plane. When the image size of the projection plane is set to H_p and W_p, the following equations are obtained for H_p and W_p:

$$H_p = 2r \tan \left(\frac{B_h}{H}\frac{\pi}{2}\right) \quad (3)$$

$$W_p = 2r \tan \left(\frac{B_w}{W}\pi\right) \quad (4)$$

TABLE 2. Selection rate for each image group.

	Random	User annotation	Our results	Tie
a=0	58/405	119/405**	**218**/405**	10/405
a=0.01	48/405	124/405**	**215**/405**	18/405
a=0.03	43/405	110/405**	**236**/405**	16/405
a=0.1	51/405	115/405**	**229**/405**	10/405
a=0.4	48/405	157/405*	**191**/405*	9/405
a=1	68/405	159/405	**164**/405	14/405

*:p<0.1,**:p<0.01

where r is the radius of the unit sphere, H and W are the height and width in pixels of the ERP image, respectively. $B_h \in (0, H)$ and $B_w \in (0, \frac{W}{2})$ are the height and width of each RoI, respectively. We extracted five perspective projection images from a 360° image based on the resulting RoIs. This resulted in 135 pairs of a ERP image and a set of five RoIs for evaluation.

For the user study, we recruited the subjects on Amazon Mechanical Turk [13] to view each 360° image in a browser for 30s, and answer a questionnaire. The protocol is that after showing the 360° images to the cloud workers for 30s, we showed them three groups of images:

A Five perspective RoI images of our user-annotated RoIs dataset.

B Five perspective RoI images predicted by the proposed method.

C Five randomly chosen regions from candidates obtained by Selective Search.

We created the following question: "Which of the three sets of regions do you think is the most impressive in the 360° image?" We also provided a choice "tie." Workers were not told which group each set of images corresponded to, and the order of the options was changed each time. We received three responses for each of the 135 sets and repeated these experiments six times with different Sailent-IoU parameters a while maintaining $n = 5$. We only approved workers who spent at least 60 seconds per one image. We did not limit the maximum number of tasks that one worker could participate in, and thus the number of workers differed from 42 to 67 as the value of a changed. In each 360° image, we got answers from different three workers.

The selection rates for each image group are shown in Table 2. Interestingly, the RoI predicted by the proposed method was rated higher than the user-annotated RoI, that is, manual choice. According to the results of χ^2 tests, on the condition of $a = 0, 0.01, 0.03, 0.1, 0.4$, there is statistical significance between the frequencies of User annotation and Our results. In addition, the RoI detected with weight on both saliency and IoU was more preferred, and it shows SIoU is the reasonable evaluation function for RoI detection.

V. CONCLUSION

In this study, we tackle the problem of predicting the regions of interest (RoI) from a single 360° image as a set of perspective projection images with variable FoV and free positioning relationships. We proposed an algorithm to predict the optimal RoI set based on the saliency map predicted from 360° images and an evaluation function considering both the saliency value and the IoU in candidate regions obtained through Selective Search. To train the network to predict the saliency map, we proposed the random rotation data augmentation to overcome the strong center bias of the training data, and showed a significant improvement in performance over the baseline [8]. We created RoI dataset, and evaluated the predected RoIs in quantitatively, qualitatively with it. In user study, we show that our algorithm can predict reasonable RoIs.

REFERENCES

[1] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2Vid: Automatic cinematography for watching 360° videos," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 154–171.

[2] B. Xiong and K. Grauman, "Snap angle prediction for 360° panoramas," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–18.

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[4] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Ann. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, 1995.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[6] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4789–4798.

[7] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "SalGAN360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2018, pp. 1–4.

[8] D. Martin, A. Serrano, and B. Masia, "Panoramic convolutions for 360° single-image saliency prediction," in *Proc. CVPR Workshop Comput. Vis. Augmented Virtual Reality*, 2020, pp. 1–4.

[9] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "SalNet360: Saliency maps for omni-directional images with CNN," *Signal Process., Image Commun.*, vol. 69, pp. 26–34, Nov. 2018.

[10] H. Lv, Q. Yang, C. Li, W. Dai, J. Zou, and H. Xiong, "SalGCN: Saliency prediction for 360-degree images based on spherical graph convolutional networks," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 682–690.

[11] J. Gutierrez, E. J. David, A. Coutrot, M. P. Da Silva, and P. L. Callet, "Introducing UN salient360! benchmark: A platform for evaluating visual attention models for 360° contents," in *Proc. 10th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2018, pp. 1–3.

[12] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proc. 8th ACM Multimedia Syst. Conf.*, Jun. 2017, pp. 205–210.

[13] (2022). *Amazon Mechanical Turk*. [Online]. Available: https://www.mturk.com/

[14] Y.-C. Su and K. Grauman, "Making 360° video watchable in 2D: Learning videography for click free viewing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1368–1376.

[15] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360° sports videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1396–1405.

[16] S. Lee, J. Sung, Y. Yu, and G. Kim, "A memory network approach for story-based temporal summarization of 360° videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1410–1419.

[17] M. Wang, Y.-J. Li, W.-X. Zhang, C. Richardt, and S.-M. Hu, "Transitioning360: Content-aware NFoV virtual camera paths for 360° video playback," in *Proc. Int. Symp. Mixed Augmented Reality (ISMAR)*, Nov. 2020, pp. 185–194.

[18] A. Bur, A. Tapus, N. Ouerhani, R. Siegwar, and H. Hiigli, "Robot navigation by panoramic vision and attention guided fetaures," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 695–698.

[19] Y. Rai, J. Gutiérrez, and P. L. Callet, "A dataset of head and eye movements for 360 degree images," in *Proc. 8th ACM Multimedia Syst. Conf.*, Jun. 2017, pp. 205–210.

[20] E. David, J. Gutiérrez, A. Coutrot, M. P. D. Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proc. 9th ACM Multimedia Syst. Conf.*, Jun. 2018, pp. 432–437.

[21] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360° videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1420–1429.

[22] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360° videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 488–503.

[23] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. G. I. Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," 2018, *arXiv:1701.01081*.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[26] H. T. T. Tran, C. T. Pham, N. P. Ngoc, A. T. Pham, and T. C. Thang, "A study on quality metrics for 360 video communications," *IEICE Trans. Inf. Syst.*, vol. E101.D, no. 1, pp. 28–36, 2018.

[27] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.

[28] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, 1999.

[29] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360° videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 488–503.

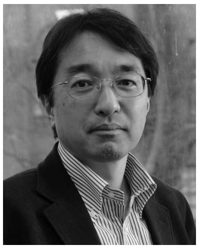
[30] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[31] Y. Zhu, G. Zhai, X. Min, and J. Zhou, "The prediction of saliency map for head and eye movements in 360 degree images," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2331–2344, Sep. 2020.

[32] (2022). *YouTube*. [Online]. Available: https://www.youtube.com/

YUUKI SAWABE (Graduate Student Member, IEEE) received the B.S. degree in electrical and electronic engineering from The University of Tokyo, Tokyo, Japan, in 2021, where he is currently pursuing the M.S. degree in information science and technology with the Graduate School of Information Science and Technology. His research interests include computer vision, 360° image processing, and VR/AR.

SATOSHI IKEHATA (Member, IEEE) received the B.A. degree in psychology, and the M.S. and Ph.D. degrees in information studies from The University of Tokyo, in 2009, 2011, and 2014, respectively. He worked as a Postdoctoral Researcher with Washington University in St. Louis, from 2014 to 2016. He is currently an Assistant Professor with the National Institute of Informatics. His research interests include 3-D computer vision, physics-based 3-D reconstruction, VR/AR, indoor/outdoor scene understanding, and human 3-D cognition and perception.



KIYOHARU AIZAWA (Fellow, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from The University of Tokyo, in 1983, 1985, and 1988, respectively. He is currently a Professor with the Department of Information and Communication Engineering, The University of Tokyo. He was a Visiting Assistant Professor with the University of Illinois, from 1990 to 1992. His research interests include multimedia applications, image processing, and computer vision. He is a fellow of IEICE and ITE and a Council Member of Science Council of Japan. He received the 1987 Young Engineer Award and the 1990, 1998 Best Paper Awards, the 1991 Achievement Award, 1999 Electronics Society Award from IEICE Japan, the 1998 Fujio Frontier Award, the 2002 and 2009 Best Paper Award, and the 2013 and 2020 Achievement Award from ITE Japan. He received the IBM Japan Science Prize in 2002. He is on Editorial Boards of IEEE MultiMedia and *ACM TOMM*. He served as the Editor-in-Chief for *ITE Journal* in Japan, and an Associate Editor for IEEE Transaction on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Multimedia. He was the President of ITE and ISS Society of IEICE, in 2019 and 2018, respectively. He has served a number of international and domestic conferences; he was the General co-Chair of ACM Multimedia 2012 and ACM ICMR2018.