Distortion-Adaptive Salient Object Detection in 360° Omnidirectional Images

Jia Li, *Senior Member, IEEE*, Jinming Su, Changqun Xia, and Yonghong Tian, *Senior Member, IEEE*

***Abstract*—Image-based salient object detection (SOD) has been extensively explored in the past decades. However, SOD on 360° omnidirectional images is less studied owing to the lack of datasets with pixel-level annotations. Toward this end, this paper proposes a 360° image-based SOD dataset that contains 500 high-resolution equirectangular images. We collect the representative equirectangular images from five mainstream 360° video datasets and manually annotate all objects and regions over these images with precise masks with a free-viewpoint way. To the best of our knowledge, it is the first public available dataset for salient object detection on 360° scenes. By observing this dataset, we find that distortion from projection, large-scale complex scene and small salient objects are the most prominent characteristics. Inspired by the founding, this paper proposes a baseline model for SOD on equirectangular images. In the proposed approach, we construct a distortion-adaptive module to deal with the distortion caused by the equirectangular projection. In addition, a multi-scale contextual integration block is introduced to perceive and distinguish the rich scenes and objects in omnidirectional scenes. The whole network is organized in a progressively manner with deep supervision. Experimental results show the proposed baseline approach outperforms the top-performanced state-of-the-art methods on 360° SOD dataset. Moreover, benchmarking results of the proposed baseline approach and other methods on 360° SOD dataset show the proposed dataset is very challenging, which also validate the usefulness of the proposed dataset and approach to boost the development of SOD on 360° omnidirectional scenes.**

***Index Terms*—Salient object detection, 360° omnidirectional image, distortion-adaptive, benchmarking.**

I. Introduction

THE purpose of image-based salient object detection (SOD) is to detect and segment objects that capture human visual attention, which is an important preliminary step for various visual tasks such as object recognition [1], tracking [2] and image parsing [3]. In the past decades, many benchmark datasets [4]–[9] have been constructed to drive the development of SOD methods. On these datasets, many learning-based methods [9]–[12] have been proposed to boot the performance of SOD and have made great progress. There is a fact that almost all existing methods focus on image-/video-based SOD, where the image/video always is displayed with a limited field of view (FoV). However, what human beings perceive is in a three-dimensional world, and the objects analyzed by human vision are 360° omnidirectional scenes at every moment. With the development of imaging technology and hardware, 360° content becomes more and more widespread on popular image/video sharing platforms. How to analyze 360° content is an emerging problem, which is important to understand and compress image/video information [13] as well as enhance user experience. As an important pre-step of most visual tasks, saliency detection is a good tool for 360° image/video analysis.

In recent years, some omnidirectional video datasets [14]–[22] have appeared. These datasets contain many 360° omnidirectional videos, and are usually used to explore the visual attention mechanism of human beings, mainly human fixation. SOD as a related task of fixation estimation, can better perceive and detect saliency in object-level, and have great significance for higher-level tasks (*e.g.* object tracking). However, there still lack 360° image/video datasets for SOD, which prevents the fast grown of this branch. Judging from the actual needs, in order to meet the needs of analyzing the growing omnidirectional imaging data by means of computer technology, it is necessary to construct a 360° datasets for SOD to promote the development of SOD on omnidirectional data.

Toward this end, this paper proposes **360-SOD**, a 360° omnidirectional image dataset for SOD, which contains 500 images with pixel-level annotation including various scenes as some representative examples shown in Fig. 1. In constructing 360-SOD, we first collect a large number of 360° videos, and segment them to get 6870 key frames. Given these images, two volunteers are asked to judge whether there contain unambiguous salient objects and whether this frame is a redundancy of adjacent image, and finally 500 images with obvious salient objects are picked out. In the next stage, six engineers are asked to manually label the accurate boundaries of salient objects. The judgment

Manuscript received April 21, 2019; revised September 6, 2019; accepted November 19, 2019. Date of publication December 6, 2019; date of current version February 5, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002400, in part by the National Natural Science Foundation of China under Contract 61672072, Contract 61922006, Contract 61825101, and Contract 61532003, and in part by the Beijing Nova Program under Grant Z181100006218063. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Ali Borji. *(Corresponding authors: Changqun Xia; Yonghong Tian.)*

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, with the Beijing Adavanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: jiali@buaa.edu.cn).

J. Su is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: sujm@buaa.edu.cn).

C. Xia is with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: xiacq@buaa.edu.cn).

Y. Tian is with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: yhtian@pku.edu.cn).

Digital Object Identifier 10.1109/JSTSP.2019.2957982

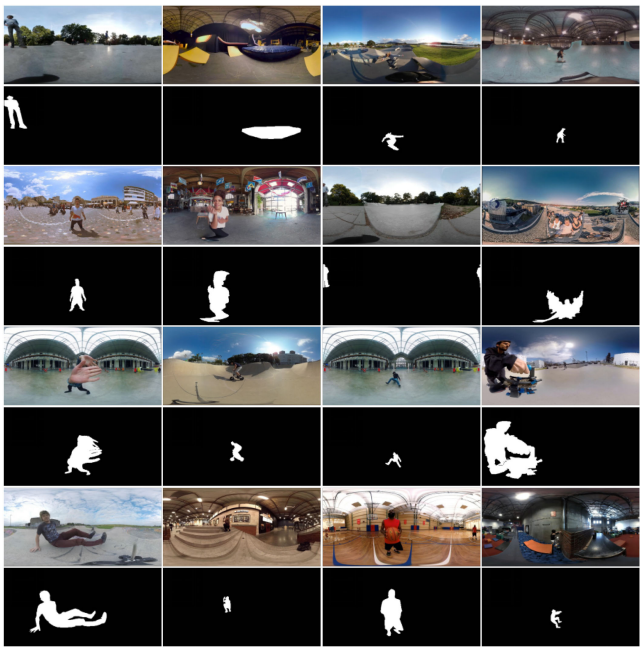

Fig. 1. Representative examples of 360-SOD. Images and ground truth are shown as equirectangular images.

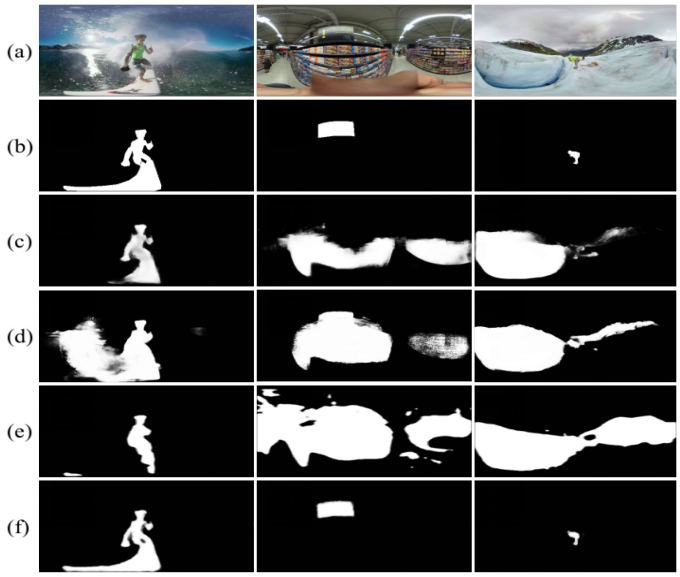

Fig. 2. Examples of predicted salient objects by different methods. (a) equirectangular form of 360° images, (b)ground truth, and saliency maps of (c) RAS [23], (d) R3Net [24], (e) DGRL [12] and (f) the proposed approach.

of salient objects is done in omnidirectional picture brower, and the annotation is based on the equirectangular form of 360° omnidirectional images. To the best of our knowledge, 360-SOD is the first public available dataset for salient object detection on 360° scenes.

Based on the omnidirectional image dataset, we explore SOD on 360° images. To predict the salient objects on 360° imaging data, an intuitive method is to store the 360° scenes as equirectangular images, and then directly utilize the SOD algorithms on conventional images. However, the predicted saliency map is not satisfactory as shown in Fig. 2. By exploring the difference between conventional images and 360° omnidirectional images, we find there are mainly three problems in the processing of omnidirectional images. The first problem is the distortion caused by projections from sphere to plane as shown in the first column of Fig. 2. No matter which projection method (*e.g.* equirectangular, cube map and patch-based projections) we choose, distortion is inevitable. In this work, we choose equirectangular projection to store and analyze these images as done in [17], [25]. The second problem is caused by large-scale complex scene in omnidirectional image as depicted in the second column of Fig. 2, which may confuse algorithms to detect wrongly. The last problem is the difficulty of perceiving and segmenting small salient objects as presented in the third column of Fig. 2. These problems make it difficult to deal with SOD on omnidirectional images.

To address these issues, we propose a baseline model on 360-SOD to consider the basic problems in omnidirectional scenes. In the proposed approach, we construct a distortion-adaptive module to deal with the distortion due to equirectangular projection. And then, the approach introduces a multi-scale contextual integration module to undertake the large-scale complex scene. Moreover, this model is organized in a progressive refinement manner to restore the boundaries and small salient objects. Experimental results show the proposed method baseline model (denoted as **DDS**) outperforms the state-of-the-art conventional SOD methods on 360-SOD.

The contributions are summarized as follows: 1) we propose the first SOD dataset on 360° omnidirectional images, which we will release to boost the development of 360° SOD, 2) we propose a baseline model to deal with the basic problems in omnidirectional scenes and it outperforms top-performanced conventional methods on 360° SOD dataset, 3) we provide a comprehensive benchmark of our approach and other state-of-the-art SOD methods on 360° scenes, which reveals the key challenges in omnidirectional scenes and validates the usefulness of the proposed dataset and baseline model.

The rest of this paper is organized as follows: Section II reviews existing datasets and models about SOD. Section III presents the new dataset on 360° omnidirectional images. In Section IV, a baseline model for SOD on 360° images is proposed. Section V benchmarks the proposed model and other state-of-the-art methods, and the paper is concluded in Section VI.

II. Related Work

SOD on 360° omnidirectional images is correlated with conventional image-/video-based SOD, and panoramic datasets and related tasks. In this section, we will review the most related datasets and models.

A. Conventional Datasets for SOD

There are many image datasets. ECSSD [4] contains 1,000 images with complex structures and obvious semantically meaningful objects. DUT-OMRON [5] consists of 5,168 complex images with pixel-wise annotations and all images are downsampled to a maximal side length of 400 pixels. PASCAL-S [6]

comprises 850 natural images that are pre-segmented into objects or regions and free-viewed by 8 subjects in eye-tracking tests for salient object annotations. HKU-IS [7] includes 4,447 images and lots of images contain multiple disconnected salient objects or salient objects that touch image boundaries. DUTS [8] is a large scale dataset containing 10,533 training images and 5,019 testing images. The images are challenging with salient objects that occupy various locations and scales as well as complex background. XPIE [9] has 10,000 images covering a variety of simple and complex scenes with salient objects of different numbers, sizes and positions. These datasets with pixel-level annotation drive the development of learning-base model [9]–[12] for image-based SOD.

Many video datasets have been proposed for SOD. SegTrack V2 [26] is a classic dataset in video object segmentation that is frequently used in many previous works. It consists of 14 densely annotated video clips with 1,066 frames in total. Youtube-Objects [27] contains a large number of Internet videos and its widely used subset [28] contains 127 videos with 20,977 frames. In these videos, 2,153 key frames are sparsely sampled and manually annotated with pixel-wise masks according to the video tags. VOS [29] contains 200 videos with 116, 093 frames. On 7,467 uniformly sampled key frames, all objects are presegmented by 4 subjects, and the fixations of another 23 subjects are collected in eye-tracking tests. These datasets provide sufficient data support for video-based SOD algorithms [30], [31].

B. Conventional Models for SOD

Hundreds of image-based SOD methods have been proposed in the past decades. The survey [32] provides detailed introduction and analysis about SOD methods, especially traditional methods. Recently, a lot of deep models are devoted to enhance the performance of neural networks for SOD. Zhang *et al.* [33] proposed an attention guided network to selectively integrates multi-level information in a progressive manner. Wang *et al.* [11] proposed a pyramid pooling module and a multi-stage refinement mechanism to gather contextual information and stage-wise results, respectively Chen *et al.* [23] proposed reverse attention mechanism which is inspired from human perception process by using top information to guide bottom-up feed-forward process in a top-down manner. Chen *et al.* [34] incorporated human fixation with semantic information to simulate the human annotation process for salient objects.

With the development of image-based SOD, video-based SOD also has made great progress. Papazoglou and Ferrari [35] proposed an approach for the fast segmentation of foreground objects from background regions. They estimated an initial foreground map with respect to the motion information, which was then refined by building the foreground/background appearance models and encouraging the spatiotemporal smoothness of foreground objects across the whole video. Wang *et al.* [36] proposed an unsupervised algorithm for video-based SOD. In their algorithm, frame-wise saliency maps were first generated and refined with respect to the geodesic distances between regions in the current frame and subsequent frames. After that, global appearance models and dynamic location models were constructed so that the spatially and temporally coherent salient objects can be segmented. Li *et al.* [31] proposes an approach for segmenting primary video objects by using Complementary Convolutional Neural Networks (CCNN) and neighborhood reversible flow. In their approach, the initialized foregroundness and backgroundness can be efficiently and accurately propagated along the temporal axis so that primary video objects gradually pop-out and distractors are well suppressed.

C. Panoramic Datasets

360-VHMD [14] is a popular dataset for 360° videos. It contains 7 videos about indoor and outdoor scenes with 48414 frames. Salient!360 [15] contains 85 equirectangular images and 19 equirectangular videos with ground-truth fixation maps and scan-paths obtained from subjective experiments. Wild-360 [18] consists of 85 360° video clips, totally about 40 k frames. 60 clips within wild-360 are for training and the rest 25 clips are for testing. All the clips are cleaned and trimmed from 45 raw videos obtained from YouTube. VR-scene [19] has 208 high definition dynamic 360° videos collected from Youtube, each with at least 4 k resolution (3840 pixels in width) and 25 frames per second. The duration of each video ranges from 20 to 60 seconds. The videos in VR-scene exhibit a large diversity in terms of contents, which include indoor scene, outdoor ac- tivities, music shows, etc. Further, some videos are captured from a fixed camera view and some are shot with a moving camera. 360-saliency [20] collects 104 video clips as the data used for saliency detection in 360° videos. The video contents involve five sports (i.e. basketball, parkour, BMX, skateboarding, and dance), and the duration of each video is between 20 and 60 seconds.

In addition, VR-VQA48 [16] is a 360° dataset for measuring the quality reduction of panoramic videos. It contains viewing direction data of 40 subjects on 48 sequences of panoramic videos. All of these sequences are downloaded from YouTube and VRCun and the resolution is beyond 3 K and up to 8 K. Sports-360 [17] consists of 342 360° videos downloaded from YouTube in five sports domains including basketball, parkour, BMX, skateboarding, and dance, which is created for the study on relieving the viewer from this "360 piloting" task. PVS-HM [21] is a new panoramic video database that consists of head movement (HM) positions across 76 panoramic video sequences with a thorough analysis. These panoramic video sequences are from YouTube and VRCun, and the duration of each sequence is cut to be from 10 to 80 seconds.

A comprehensive statistic of these datasets is listed in Table I. These datasets almost contain annotation of human fixation, while there is no annotation of salient objects. Although a large number of approaches for SOD have been proposed, there is no approach to focus on SOD on omnidirectional images/videos. In addition, even if many omnidirectional datasets have been constructed, there are few 360° datasets for the task of SOD on 360° scenes.

III. A New Dataset for 360° SOD

In this section, we will introduce the rules and details in constructing the 360° image dataset.

TABLE I
COMPARISONS BETWEEN REPRESENTATIVE 360° IMAGE/VIDEO DATASETS. #IMAGE/VIDEO: THE NUMBER OF IMAGE/VIDEO, #FRAMES: THE TOTAL NUMBER OF FRAMES AND #TIME: THE TOTAL NUMBER OF SECONDS

Dataset	Type of Scene	#Image	#Video	#Resolution(in pixels)			#Frames	#Time (in seconds)
				Width	Height	Max Resolution		
360-VHMD [14]	indoor & outdoor	-	7	[3840, 3840]	[2048, 2160]	3840 × 2160	48,414	1,340
Salient!360 [15]	indoor & outdoor	85	19	[3840, 18332]	[1920, 9166]	18332 × 9166	10,548	381
Wild-360 [18]	outdoor	-	85	[1920, 2160]	[960, 1080]	2160 × 1080	40,290	1,553
VR-scene [19]	indoor & outdoor	-	208	[1920, 7680]	[1080, 3840]	7680 × 3840	215,457	7,511
360-saliency [20]	indoor & outdoor	-	104	[3724, 3840]	[1862, 2160]	3840 × 2160	76,611	-
VR-VQA48 [16]	indoor & outdoor	-	48	[2880, 7680]	[1440, 3840]	7680 × 3840	35,906	1326
Sports-360 [17]	indoor & outdoor	-	342	[3724, 3840]	[1862, 2160]	3840 × 2160	180,000	-
PVS-HM [21]	indoor & outdoor	-	76	[2880, 7680]	[1440, 3840]	7680 × 3840	-	2,045

Fig. 3. Examples of adjacent four key frames in raw 360° images.

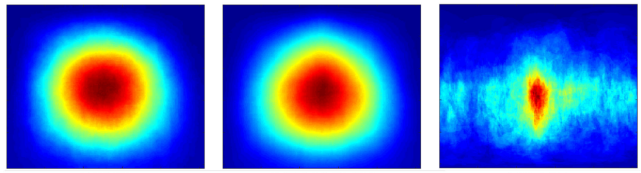

Fig. 4. Average annotation maps of two conventional image-based SOD datasets and 360-SOD.

A. Data Collection

As introduced in Section II, there exist many omnidirectional video datasets. 360-VHMD [14], Salient!360 [15], Wild-360 [18], VR-scene [19] and 360-saliency [20] are almost for the research of visual attention and contain the ground truth of human fixation. Therefore, there datasets are likely to have regions or objects that can attract human visual attention. Instead of undertaking a new data shooting and video recording, we directly combine these five datasets as a raw data source. Next, we intercept a key frame every two seconds from these videos in the data source. We collect these key frames and merge them with the original images. After that, we resize each image to have a maximum side length of 1024 pixels for convenient processing. Finally, we obtain a raw 360° omnidirectional image dataset with 6870 images. Note that because the 360-saliency doesn't provide the video frame rate, we adopt 30FPS as its default value.

B. Annotation of Salient Objects

There exist some scenes with unclear salient objects and redundant data between continuous key frames in above raw 360° image dataset, as shown in Fig. 3. Therefore, we divide the process of annotation of this dataset into two stages. In the first stage, these 6870 images are displayed in omnidirectional picture brower in original order. Then, we ask two volunteers to judge whether an image is selected according to three main principles, which include: (1) there is clear scene without jitter, (2) unambiguous, meaningful and annotatable salient objects exist in the scene, and (3) if the adjacent multiple images has same salient objects with similar shape and position, only the best one is selected. If the answer is "Yes," we will collect this image. After the first stage that 6870 360° images are processed, we finally collect 500 images from the raw 360° image dataset. In the second stage, six engineers are asked to judge the salient objects in omnidirectional picture brower and manually label the accurate boundaries of salient objects by LabelMe Annotation Tool [37] in the equirectangular form of these 500 images. The manual annotation is a time-consuming work, and it takes an average of five minutes to annotate each image. Note that we have two volunteers involved in the process for cross-check the quality of annotations.

Finally, we obtain the final 360° image dataset, which contains 500 equirectangular images with pixel-level annotation. This dataset includes indoor and outdoor scenes from different perspectives as well as different scene complexities. The dataset is denoted as **360-SOD**. Some representative examples of 360-SOD can be found in Fig. 1.

C. Dataset Statistics

To explore the main characteristics of 360-SOD, we present the average annotation maps (AAMs) of 360-SOD and two conventional image-based SOD datasets (*i.e.*, ECSSD [4] and DUT-OMRON [5]) as shown in Fig. 4. As in [32], the AAM of an image-based SOD dataset is computed by 1) resizing all ground-truth masks from the dataset to the same resolution, 2) summing the resized masks pixel by pixel, and 3) normalizing the resulting map to a maximum value of 1.0. In this way, the figure gives a better view of the distribution of salient objects in all the images in 360-SOD.

From Fig. 4, we can see that the distributions of conventional image-based SOD datasets are usually center-biased, while the distribution of the 360° omnidirectional image-based dataset 360-SOD is more discrete. This may be caused by the characteristic of large-scale complex scenes in 360° image, which indicates the diversity of 360-SOD and the difficulty of 360° SOD.

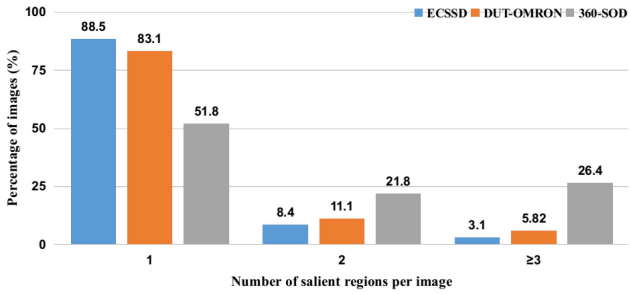

Fig. 5. Histograms of the number of salient objects in 3 datasets.

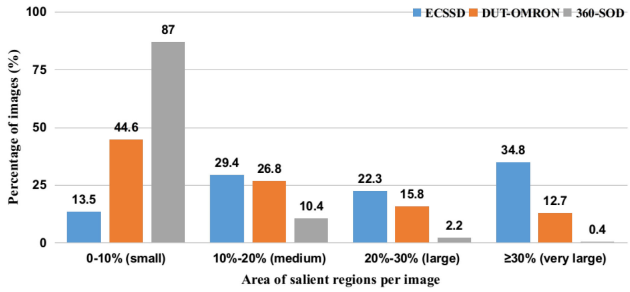

Fig. 6. Histograms of the area of salient objects in 3 datasets.

In addition, we present the histograms of number and area of salient objects in 360-SOD and other two datasets in Fig. 5 and Fig. 6. As shown in Fig. 5, we can find that there are usually more salient objects in 360-SOD that other two datasets, which is caused by the difference of conventional and 360° images. Moreover, we can see there are more images with small areas of salient objects in Fig. 6, which maybe one of the main challenges in 360° omnidirectional images.

D. Analysis

We directly deal with this dataset by existing conventional SOD algorithm and the results are presented in Fig. 2, which are obviously not satisfactory. We explore the difference between conventional images and 360° omnidirectional images, and we find there are three main problems leading to the difficulty in 360° SOD: 1) distortion, which is inevitable because of the projection from sphere to plane; 2) large-scale complex scene, which is the main characteristics of omnidirectional images; and 3)small salient objects that commonly exist in 360° images as presented in Fig. 2. These issues actually prevent the detection of salient objects in 360° images.

IV. A Baseline Model of SOD on 360° Images

To deal with the three existing issues in omnidirectional images (*i.e.*distortion from projection, large-scale complex scene and small salient objects), we propose a baseline model for SOD on 360° images. The baseline model is inspired by the existing issues, and it is fed by the equirectangular image as input and outputs a saliency map with the same resolution as input as presented in Fig. 7. Details of the proposed approach are descried as follows.

A. Architecture

As depicted in Fig. 1, the baseline model is a distortion-adaptive network with deep supervision (denoted as **DDS**). The first module of DDS is a distortion-adaptive module, which is designed to deal with the distortion caused by the projections from sphere to plane. After the processing of distortion-adaptive module, the distortion in the image will be corrected adaptively and a new image is output. Following this module, DDS takes ResNet-50 [38] as the feature extractor, which is modified to remove the last global pooling and fully connected layers for the pixel-level prediction. Feature extractor has five residual modules, named as $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_5$. To obtain larger feature maps, the strides of all convolutional layers belonging to last two residual modules $\mathcal{R}_4$ and $\mathcal{R}_5$ are set to 1. To further enlarge the receptive fields of high-level features, we set the dilation rates [39] to 2 and 4 for convolutional layers in $\mathcal{R}_4$ and $\mathcal{R}_5$, respectively. For a $H \times W$ input images, a $\frac{H}{8} \times \frac{W}{8}$ feature map is output by the feature extractor.

We select the outputs of last convolutional layers in $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_5$ as side-outputs as used in [23], [40]. For each side-output, its features are processed by a convolutional layer with kernels 1×1 to compress the feature channel. For the side-output of $\mathcal{R}_5$, after the channel compression layer, a multi-scale context integration module is conducted to integrate multi-scale contextual information, which is for undertaking the large-scale complex scene. The multi-scale context integration module is a variant of Atrous Spatial Pyramid Pooling (ASPP) [39] as shown in Fig. 8. The module is composed of four branches with four dilation convolutional layers. The four layers consist of 128 kernels of 3×3 with dilation rates of 1, 2, 3 and 4, followed by an element-wise summation operation. The multi-scale context integration module is able to effectively integrate multi-scale features. Following this module, another convolutional layer with one kernel 1×1 to convert the feature space from high dimension to saliency features with one channel. Moreover, these saliency features are upsampled and concatenated into the output of $\mathcal{R}_4$ to get finer saliency features. This mechanism of that coarser-level saliency features are concatenated into the adjacent finer-level features are conducted on every two adjacent side-outputs, and the whole network is organized in a progressive form.

B. Distortion Adaptation for Equirectangular Images

To deal with the projection distortion in equirectangular images, a distortion-adaptive module is conducted to correct the distortion of the input images adaptively. The equirectangular projection is one of most commonly used projection methods from sphere to plane, which will bring different degrees of distortion in different positions of the image, especially locations near the poles.

To address this distortion, we propose a distortion-adaptive module to deal with different regions with various parameters. As shown in Fig. 9(a), we cut the input equirectangular image into $N \times N$ image blocks and we use I_{ij} $(i, j = 1, 2, \ldots, N)$ that represents the image block at ith row and jth column, and K_{ij} which represents the corresponding kernels of I_{ij}. In this

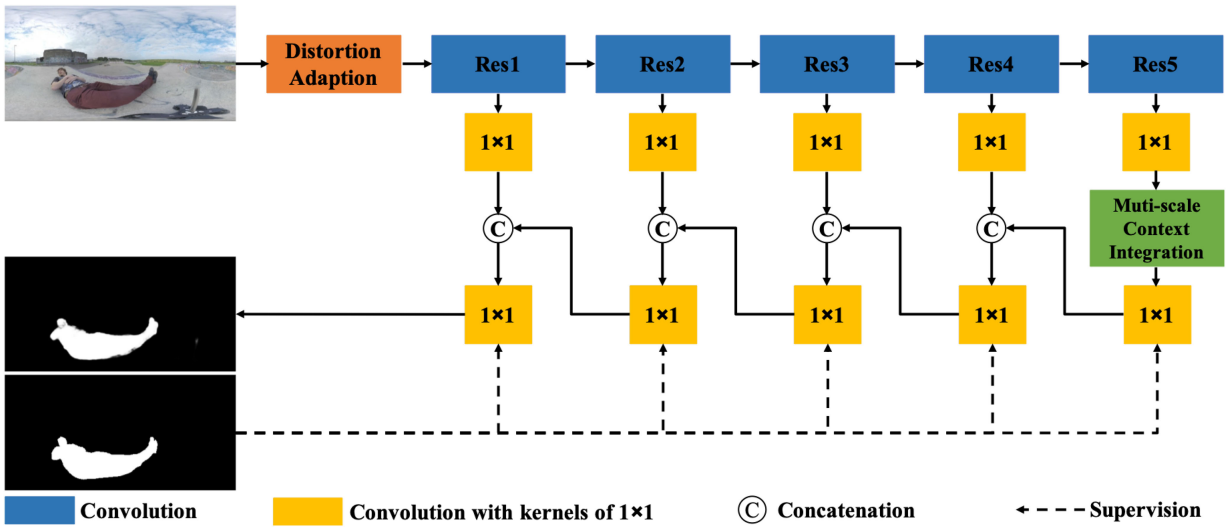

Fig. 7. The framework of our baseline model. The equirectangular image is processed by a distortion-adaptive module, and the output is transferred to ResNet-50 to extract features in different levels. The highest-level features are dealt with by a multi-scale context integration module to integrate information. Moreover, the coarser-level saliency features are concatenated into the adjacent finer-level features to get finer saliency maps and the whole network is organized in a progressive form.

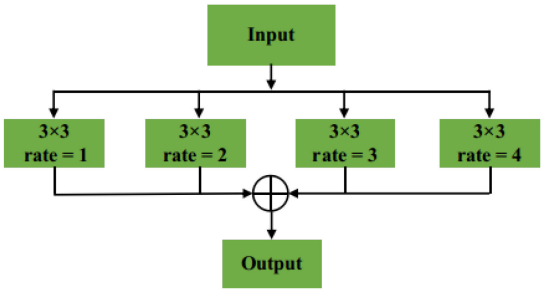

Fig. 8. Structure of the multi-scale context integration module.

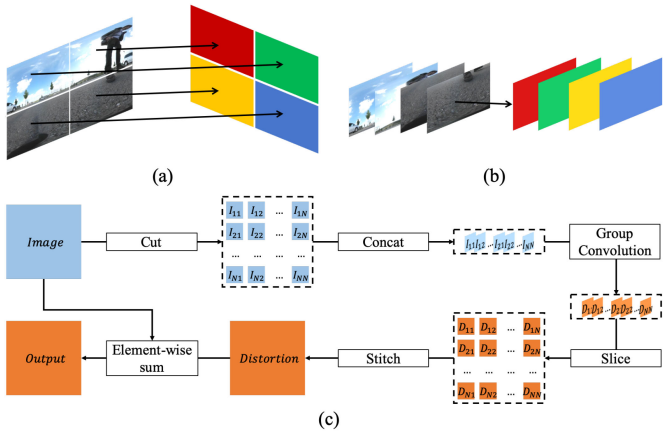

Fig. 9. Details of distortion-adaptive module. In this example, the image is cut into 4 image blocks, and the 4 blocks are convolved with 4 different convolutional kernels that are display as 4 colors in this figure. The convolutional kernels are learnable. Note we choose the number of image blocks as 4 for convenient display. (a) and (b) are equivalent forms, and (c) is the complete structure of distortion-adaptive module, where I_{ij}, D_{ij}, N and other details are introduced in Section VI-B.

manner, I_{ij} is only convolved with the K_{ij} and K_{ij} is only used to convolve I_{ij}. Moreover, K_{ij} has three channels and I_{ij} has three kernels with three channels. In this module, different image blocks are convolved with different kernels. The output has the same resolution as input, and it also can be regarded as $N \times N$ image blocks, which is the learned distortion. We denote the distortion at ith row and jth column of output as D_{ij}, so D_{ij} is computed by

$$D_{ij} = I_{ij} * K_{ij}, \tag{1}$$

where $*$ mean the standard convolution operation.

In a specific implementation, we can conduct this operation by group convolution operation as presented in Fig. 9(b). In detail, we first cut the input image into $N \times N$ image blocks with 3 channels, and then we concatenate these image blocks in the dimension of channel. Next, a group convolutional layer with $N \times N$ kernels is conducted, which is equivalent to directly convolve different image blocks with different kernels. The last step is to slice the learned distortion as the reverse operation of concatenation as well as stitch image blocks in the reverse manner of cutting operation. In particular, we organize the distortion-adaptive module in residual learning whose complete structure is depicted in Fig. 9(c). O_{ij} in output can be computed by

$$O_{ij} = I_{ij} + I_{ij} * K_{ij}. \tag{2}$$

Then, this output is further encoded and decoded by the following neural network, and finally supervised by the ground truth of salient objects. In this process, all parameters in the whole network (including parameters in the convolutional layer of the distortion-adaptive module) are constrained by the supervisory signal of the ground truth of salient objects, so as to realize parameter learning through gradient back-propagation. Therefore, through end-to-end training of the model, the distortion-adaptive module can be supervised and learned.

TABLE II
PERFORMANCE BENCHMARK OF 13 STATE-OF-THE-ART MODELS BEFORE BEING FINE-TUNED ON 360-SOD. THE BEST THREE RESULTS ARE IN RED, GREEN AND BLUE

	ELD [41]	UCF [42]	NLDF [43]	Amulet [44]	FSN [34]	SRM [11]	C2SNet [45]	RAS [23]	PiCANet [46]	R3Net [24]	DGRL [12]	RFCN [47]	DSS [40]
MAE ↓	0.135	0.237	0.089	0.191	0.115	0.123	0.144	0.079	0.133	0.101	0.135	0.103	0.094
F_β^w ↑	0.213	0.203	0.339	0.226	0.289	0.302	0.266	0.395	0.327	0.341	0.342	0.270	0.338
F_β ↑	0.234	0.248	0.369	0.260	0.321	0.357	0.290	0.417	0.363	0.408	0.415	0.325	0.356

C. Supervision

We adopt deep supervision applied to each side-output as used in [23], [40]. In our network, each side-output is penalized by a standard binary cross-entropy loss. Formally, we denote the cross-entropy loss function of sth side-output as $\mathcal{L}^{(s)}$, which is computed by the following formulation:

$$\begin{aligned}&\mathcal{L}^{(s)}(\mathbf{I}, \mathbf{G}, \mathbf{W}, \mathbf{w}^{(s)})\\ &= -\sum_{l=1}^{|\mathbf{I}|} \mathbf{G}(l)\log(P(\mathbf{G}(l)=1|\mathbf{I}(l);\mathbf{W},\mathbf{w}^{(s)}))\\ &\quad + (1-\mathbf{G}(l))\log(P(\mathbf{G}(l)=0|\mathbf{I}(l);\mathbf{W},\mathbf{w}^{(s)})),\end{aligned} \tag{3}$$

where $\mathbf{I}$ and $\mathbf{G}$ represent the input image and the corresponding ground truth, and $\mathbf{W}$ denotes the set of parameters of the feature extractor while $\mathbf{w}^{(s)}$ refers to the parameters of all the layers at sth side-output. In addition, $P(\mathbf{G}(l)=1|\mathbf{I}(l);\mathbf{W},\mathbf{w}^{(s)})$ represents the probability of the activation value at location l in the sth side output where l is the spatial coordinate.

Then, the overall learning objective can be formulated as:

$$\min_{\mathbf{W},\mathbf{w}} \sum_{s=1}^{S} \mathcal{L}^{(s)}(\mathbf{I}, \mathbf{G}, \mathbf{W}, fw^{(s)}), \tag{4}$$

where S is the total side-output number, and $\mathbf{w}$ is the collections of parameters of all layers at all side-outputs, which is represented by

$$\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots, \mathbf{w}^{(S)}). \tag{5}$$

In this work, S is equal to 5.

V. EXPERIMENTS

In this section, we benchmark the proposed approach DDS and other 13 state-of-the-art SOD methods on the proposed 360° dataset 360-SOD.

A. Experimental Settings

1) Dataset: In the comparisons, we divide 360-SOD into two subsets: 80% for training (400 images) and 20% for testing (100 images) by random shuffle algorithm.

2) Evaluation Metrics: We adopt mean absolute error (MAE), F-measure score (F_β) and weighted F-measure score (F_β^w) [48] as our evaluation metrics. MAE reflects the average pixel-wise absolute difference between the estimated and ground-truth saliency maps. In computing F_β, we normalize the predicted saliency maps into the range of [0, 255] and binarize the saliency maps with a threshold sliding from 0 to 255 to compare the binary maps with ground-truth maps. At each threshold, Precision and Recall can be computed. F_β is computed as:

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision}\cdot \text{Recall}}{\beta^2\cdot \text{Precision} + \text{Recall}}. \tag{6}$$

where we set $\beta^2 = 0.3$ to emphasize more on Precision than Recall as suggested in [49]. We report F_β using an adaptive threshold for generating binary a saliency map and the threshold is computed as twice the mean of a saliency map. Besides, F_β^w is computed to reflect the overall performance (refer to [48] for details).

3) Training and Inference: We use standard stochastic gradient descent algorithm to train the whole network end-to-end with the cross-entropy losses between estimated saliency maps and ground-truth masks. In the optimization process, the parameter of feature extractor is initialized by the pre-trained ResNet-50 model [38], whose learning rate is set to 5×10^{-9} with a weight decay of 0.0005 and momentum of 0.9. The learning rates of the rest layers in our network are set to 10 times larger. Besides, we employ the "poly" learning rate policy for all experiments similar to [39].

We train our network by utilizing the training set of 360-SOD, which comprises of per-pixel ground-truth annotation for 400 images. The training images are resized to the resolution of 512×256 with the treatment of horizontal flipping. In our experiment, we cut the input image into 4×4 image blocks. The training process takes about 1.5 hours and converges after 50 k iterations with mini-batch of size 1 on a single GTX 1080ti GPU. During testing, the proposed network removes all the losses, and each image is directly fed into the network to obtain its saliency map at the first side-output without any pre-processing. Due to the limitation of GPU memory and to improve the training efficiency, we downsample to a maximum side length of 512 pixels to conduct all experiments.

B. Model Benchmarking

To show the challenges of 360-SOD, we list the state-of-the-art model performance in Table II before fine-tuning them on 360-SOD. These models include ELD [41], UCF [42], NLDF [43], Amulet [44], FSN [34], SRM [11], C2SNet [45], RAS [23], PiCANet [46], R3Net [24], DGRL [12], RFCN [47] and DSS [40].

On this dataset, DSS and RAS achieve the good performance, which maybe indicates the deep supervision is beneficial to the 360° image-based SOD. Moreover, R3Net and DGRL also obtain good results. R3Net is a method that combines high-level and low-level features, and dense conditional random field (dense CRF) [50] is adopted for post-processing. DGRL is proposed to learn the local contextual information for each spatial

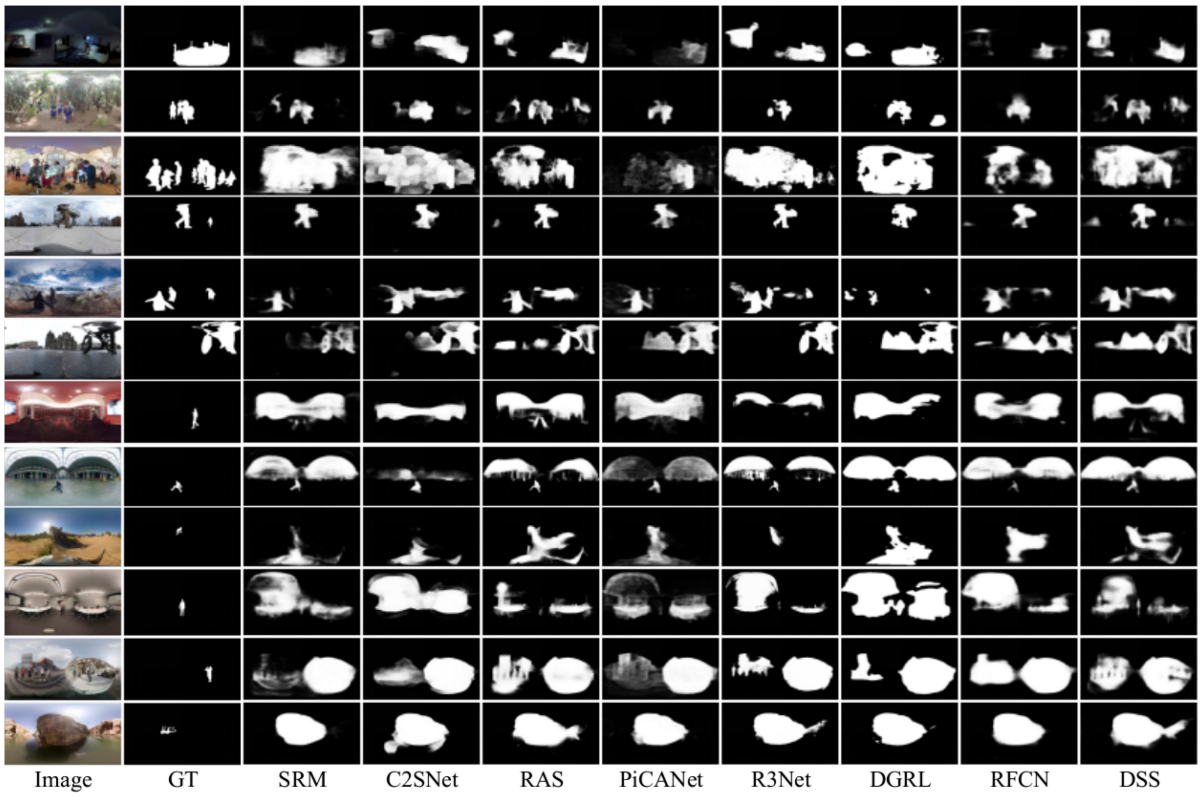

Fig. 10. Representative examples of the state-of-the-art algorithms on 360-SOD without fine-tuning.

position to refine boundaries. From these comparisons, we can believe the different-level features, different-scale contextual information and boundary refinement may be useful for saliency detection on omnidirectional image.

Moreover, some representative examples are shown in Fig. 10 for comprehensive analysis of these state-of-the-art on 360-SOD. The figure shows best and worst examples of 8 state-of-the-art conventional SOD methods. We can observe that the large-scale salient objects in simple scene usually can be detected by all these approaches, while the small salient objects in complex scene will cause all approaches to fail.

Beyond the direct performance comparisons without fine-tuning, we fine-tune the proposed method and the state-of-the-art models on the training set of 360-SOD dataset. The performance scores of the fine-tuned models on 360-SOD are listed in Table III. Some representative results of DDS and other methods are shown in Fig. 11. Comparing Table II and Table III, we can observe that the performance of all the methods is all improved. In Table III, it is worth noting that F_β^w of our method is significantly better compared with the second best results (0.652 against 0.591). Our method also consistently outperforms other models on MAE and F_β except R3Net with dense CRF [50]. The dense CRF post-processes the predicted saliency maps to obtain sharper boundaries, which will improve performance, but also greatly increase the inference time. In Fig. 11, we can find the proposed method has better results comparing with other state-of-the-art methods on 360-SOD.

From these experiments, we can believe that the proposed 360° image dataset 360-SOD is a challenging dataset for omnidirectional scenes. Moreover, from the comparisons of the proposed baseline model and other state-of-the-art methods, we can know the proposed model has good performance and be useful for solving the problem of SOD.

TABLE III
PERFORMANCE OF DDS AND THE STATE-OF-THE-ART MODELS AFTER BEING FINE-TUNED ON 360-SOD. NOTE "R3Net-w/oCRF" MEANS R3Net WITHOUT DENSE CRF [50] AND "-" MEANS THE TRAINING CODE IS NOT AVAILABLE. THE BEST THREE RESULTS ARE IN RED, GREEN AND BLUE

	MAE ↓	F_β^w ↑	F_β ↑
ELD [41]	-	-	-
UCF [42]	0.046	0.353	0.372
NLDF [43]	0.042	0.402	0.424
Amulet [44]	0.031	0.526	0.583
FSN [34]	0.030	0.529	0.609
SRM [11]	0.028	0.538	0.593
C2SNet [45]	-	-	-
RAS [23]	0.031	0.555	0.537
PiCANet [46]	**0.026**	0.578	0.589
R3Net-w/oCRF [24]	0.029	0.546	0.599
R3Net [24]	0.028	0.551	**0.677**
DGRL [12]	0.042	0.427	**0.630**
RFCN [47]	0.027	**0.585**	0.603
DSS [40]	**0.025**	**0.591**	0.599
Ours	**0.023**	**0.652**	**0.650**

C. *Performance Analysis of the Baseline Model*

1) Generalization Ability: To validate the generalization ability of the proposed baseline model on 360-SOD, we randomly

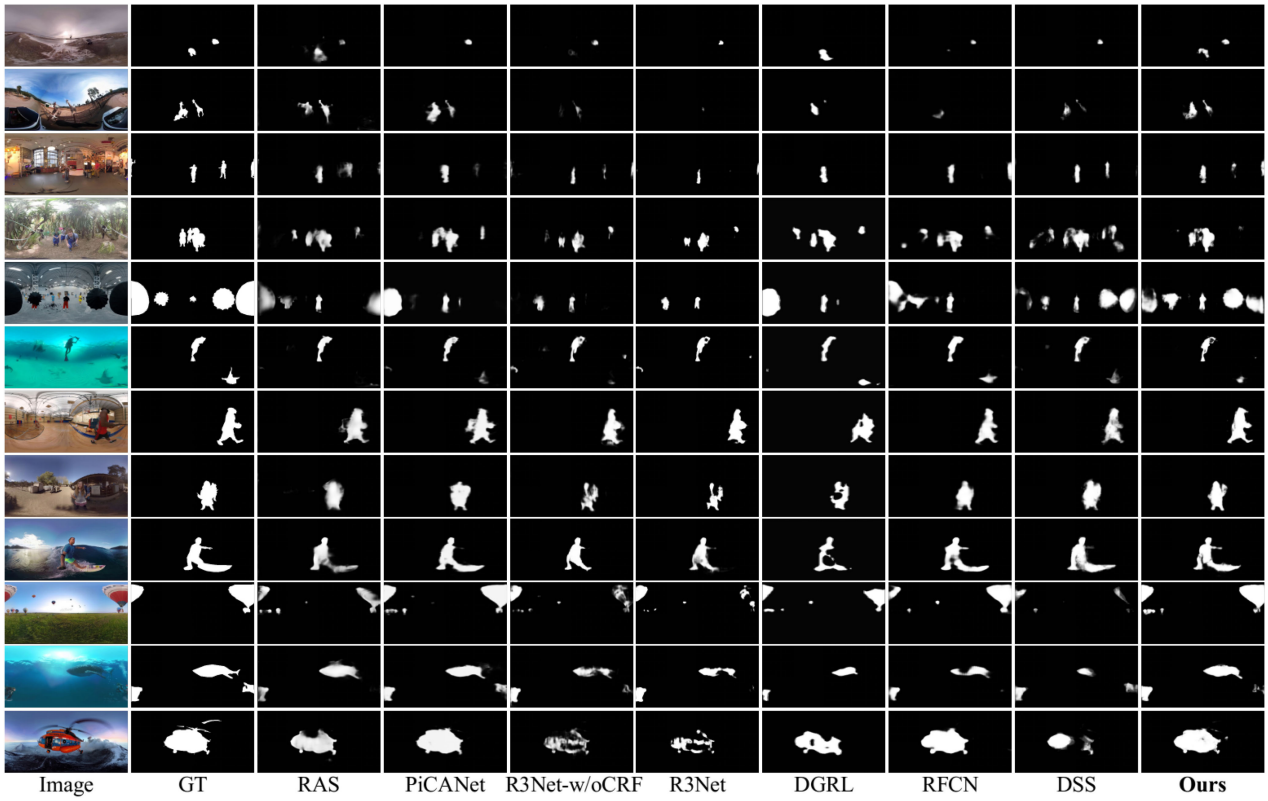

Fig. 11. Representative examples of the state-of-the-art algorithms after being fine-tuned on 360-SOD. "R3Net-wo" means R3Net without dense CRF [50].

TABLE IV
PERFORMANCE OF DDS ON RANDOMLY DIVIDED 360-SOD FOR THREE MORE TIMES

	MAE ↓	F_β^w ↑	F_β ↑
1	0.022	0.665	0.651
2	0.023	0.664	0.637
3	0.023	0.654	0.644

TABLE V
PERFORMANCE OF DDS AND THE STATE-OF-THE-ART MODELS ON 360-SOD-AT. NOTE "R3Net-w/oCRF" MEANS R3Net WITHOUT DENSE CRF [50] AND "-" MEANS THE TRAINING CODE IS NOT AVAILABLE. THE BEST THREE RESULTS ARE IN RED, GREEN AND BLUE

	MAE ↓	F_β^w ↑	F_β ↑
ELD [41]	-	-	-
UCF [42]	0.047	0.352	0.361
NLDF [43]	0.041	0.408	0.419
Amulet [44]	0.031	0.508	0.581
FSN [34]	0.030	0.564	**0.620**
SRM [11]	**0.027**	0.555	0.590
C2SNet [45]	-	-	-
RAS [23]	0.030	0.574	0.544
PiCANet [46]	0.028	**0.598**	0.582
R3Net-w/oCRF [24]	0.030	0.557	0.598
R3Net [24]	0.029	0.568	**0.675**
DGRL [12]	0.043	0.429	**0.620**
RFCN [47]	0.028	0.589	0.603
DSS [40]	**0.025**	**0.627**	0.605
Ours	**0.025**	**0.656**	**0.641**

re-divide the training and testing subsets on the 360° dataset and re-train DDS. This operation is conducted three times and the experimental results are shown in Table IV. From Table IV, we can believe DDS can stably achieve good performance on omnidirectional datasets.

In order to further verify the generalization ability of DDS, we collect 50 omnidirectional images based on Sports-360 dataset [17] for annotating and construct a new testing dataset for additional testing (named as **360-SOD-AT**). It is worth noting that the distribution of the constructed dataset is relatively different from that of 360-SOD. The dataset is constructed and annotated using the process described in Section III. DDS and other state-of-the-art models are evaluated on the new testing dataset and the results are listed in Table V. From Table V, we can see that proposed method perfeorms well and consistently outperforms other state-of-the-art methods, which validates the good generalization ability of the proposed method.

2) Influence of Various Components: To validate the effectiveness of different components of the proposed method, we conduct several experiments on 360-SOD. Firstly, we construct a naive model only with the feature extractor as described in Section IV and shown in Fig. 7 and three convolutional layers, which is denoted as "FCN-dilation." Next, we construct three models that add different components to "FCN-dilation," including distortion-adaptive module (named as DA), multi-scale context integration module (named as MCI) and deep supervision (named as DS). Moreover, we combine these three components in pairs to obtain the other three models. These models are

TABLE VI
PERFORMANCE OF DIFFERENT SETTINGS ABOUT DDS ON 360-SOD. "DA" MEANS DISTORTION ADAPTATION, "MCI" REPRESENTS MULTI-SCALE CONTEXT INTEGRATION AND "DS" IS DEEP SUPERVISION. #PARAMES: THE NUMBER OF PARAMETERS (MILLIONS), #FLOPS: FLOATING POINT OPERATIONS (BILLIONS) AND #TIME: AVERAGE TESTING TIME (MILLISECOND)

	MAE ↓	F_β^w ↑	F_β ↑	#Params	#FLOPs	#Time
FCN-dilation	0.025	0.611	0.632	27.2M	59.7B	52.0ms
FCN-dilation + DA	0.025	0.623	0.637	27.2M	59.8B	53.5ms
FCN-dilation + MCI	0.025	0.614	0.635	27.0M	59.4B	53.3ms
FCN-dilation + DS	**0.023**	0.652	0.618	27.4M	60.7B	56.9ms
FCN-dilation + DA + MCI	0.024	0.624	0.642	27.0M	59.4B	54.3ms
FCN-dilation + DA + DS	**0.023**	**0.658**	0.633	27.4M	60.8B	57.7ms
FCN-dilation + MCI+ DS	0.024	0.651	0.626	27.2M	60.3B	57.5ms
DDS	**0.023**	0.652	**0.650**	27.2M	60.4B	59.0ms

trained on the training set of 360-SOD and the performance on the testing set of 360-SOD is shown in Table VI. From Table VI, we can find deep supervision is useful for the improvement of F_β^w by comparing settings with and without "DS". Moreover, it is easy to find that "DA" and "MCI" can stably promote the performance of DDS.

In addition, we analyze the parameters, floating points operations (FLOPs) and average testing time of ablation models as listed in Table VI. From Table VI, we can observe that DA boosts the performance with slight additional parameters, FLOPs and time consumption. Meanwhile, MCI reduces the amount of parameters and FLOPs due to the reduction of channels (from 512 to 128), and improves the performance, but results in a little of time consumption. Also, DS obviously improves the performance while some additional parameters, FLOPs and time are used. From the above analysis, we can see that each component can improve performance, but extra parameters and computation are introduced. In fact, this performance improvement is not caused by simple parameters or FLOPs increase and an obvious example is that MCI improves its performance with parameters and FLOPs decrease. Therefore, it can be considered that the design of each component is effective and efficient. Finally, our model integrates the three proposed components, which has obvious performance improvement compared with the naive model "FCN-dilation," but only adds a small amount of FLOPs and testing time.

VI. CONCLUSION

Due to the lack of omnidirectional datasets for salient object detection (SOD), the development of 360° SOD is restricted. To address this problem, we construct the public available 360° image-based SOD dataset **360-SOD**. This dataset contains various scenes from different perspectives as well as scene complexities. Moreover, we analyze the dataset and find three existing problems that lead to the difficulty of 360° SOD, *i.e.*distortion from projection, large-scale complex scene and small salient objects. Inspired by these problems, a baseline model is proposed on 360° SOD. This model organized in a progressive manner, utilizes the distortion-adaptive module and multi-scale context integration module to deal with existing problems. In addition, we provide a comprehensive benchmark of our approach and other 13 state-of-the-art conventional SOD algorithms on the proposed 360° SOD dataset, which shows the key challenges in omnidirectional scenes and validates the effectiveness of the proposed dataset and baseline model. We believe that the dataset and baseline model are helpful for the development of 360° SOD.

REFERENCES

[1] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2014.

[2] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 597–606.

[3] B. Lai and X. Gong, "Saliency guided dictionary learning for weakly-supervised image parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3630–3639.

[4] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1155–1162.

[5] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.

[6] 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. Pattern Recognit.*, 2014, pp. 280–287.

[7] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5455–5463.

[8] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 136–145.

[9] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4142–4150.

[10] 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. Pattern Recognit.*, 2013, pp. 2083–2090.

[11] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4019–4028.

[12] T. Wang *et al.*, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3127–3135.

[13] K.-T. Ng, S.-C. Chan, and H.-Y. Shum, "Data compression and transmission aspects of panoramic videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 82–95, Jan. 2005.

[14] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 199–204.

[15] 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.*, 2017, pp. 205–210.

[16] M. Xu, C. Li, Y. Liu, X. Deng, and J. Lu, "A subjective visual quality assessment method of panoramic videos," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2017, pp. 517–522.

[17] 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.*, 2017, pp. 1396–1405.

[18] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1420–1429.

[19] Y. Xu *et al.*, "Gaze prediction in dynamic 360° immersive videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5333–5342.

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

[21] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2693–2708, Nov. 2019.

[22] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019.

[23] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.
[24] Z. Deng *et al.*, "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 684–690.
[25] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2vid: Automatic cinematography for watching 360° videos," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 154–171.
[26] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2192–2199.
[27] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3282–3289.
[28] S. D. Jain and K. Grauman, "Supervoxel-consistent foreground propagation in video," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 656–671.
[29] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349–364, Jan. 2018.
[30] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.
[31] J. Li, A. Zheng, X. Chen, and B. Zhou, "Primary video object segmentation via complementary CNNs and neighborhood reversible flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1417–1425.
[32] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[33] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 714–722.
[34] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic CNNs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1050–1058.
[35] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1777–1784.
[36] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3395–3402.
[37] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[40] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.
[41] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 660–668.
[42] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 212–221.
[43] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6609–6617.
[44] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.
[45] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 355–370.
[46] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3089–3098.
[47] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2018.
[48] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.
[49] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.
[50] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

Jia Li (M'12–SM'15) received the B.E. degree from Tsinghua University, Beijing, China, in 2005, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2011. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing. Before he joined Beihang University in June 2014, he used to conduct research in Nanyang Technological University, Peking University, and Shanda Innovations. He is the author or co-author of more than 70 technical articles in refereed journals and conferences such as TPAMI, IJCV, TIP, CVPR, ICCV, AAAI, and ACM MM. His research interests include computer vision and multimedia big data, especially the learning-based visual content understanding. He is a senior member of CIE and CCF. More information can be found at http://cvteam.net.

Jinming Su is currently working toward the master's degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include computer vision and deep learning.

Changqun Xia received the Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China, in 2019. He is currently an Assistant Professor with Peng Cheng Laboratory, Shenzhen, China. His research interests include computer vision and image understanding.

Yonghong Tian (S'00–M'06–SM'10) is currently a Boya Distinguished Professor with the School of EECS, Peking University, Beijing, China, and is also the Deputy Director of Artificial Intelligence Research Center, Peng Cheng Laboratory, Shenzhen, China. He is the author or co-author of more than 180 technical articles in refereed journals such as IEEE TPAMI/TNNLS/TIP/TMM/TCSVT/TKDE/TPDS, ACM CSUR/TOIS/TOMM and conferences such as NeurIPS/CVPR/ICCV/AAAI/ACMMM/WWW. His research interests include computer vision, multimedia big data, and brain-inspired computation. He was/is an Associate Editor of the IEEE TCSVT (since January 2018), IEEE TMM (from August 2014 to August 2018), IEEE Multimedia Mag. (since January 2018), and IEEE Access (since January 2017). He co-initiated the IEEE International Conference on Multimedia Big Data (BigMM) and served as the TPC Co-Chair of BigMM 2015, and also served as the Technical Program Co-Chair of IEEE ICME 2015, IEEE ISM 2015, and IEEE MIPR 2018/2019, and General Co-Chair of IEEE MIPR 2020. He is the Steering Member of IEEE ICME (since 2018) and IEEE BigMM (since 2015), and is a TPC Member of more than 10 conferences, such as CVPR, ICCV, ACM KDD, AAAI, ACM MM, and ECCV. He was the recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology awards and three ministerial-level awards in China, and obtained the 2015 EURASIP Best Paper Award for Journal on Image and Video Processing, and the Best Paper Award of IEEE BigMM 2018. He is a senior member of CIE and CCF, and a member of ACM.