# View-aware Salient Object Detection for 360° Omnidirectional Image

Junjie Wu, Changqun Xia, Tianshu Yu, Jia Li, *Senior Member, IEEE*

*Abstract*—Image-based salient object detection (ISOD) in 360° scenarios is significant for understanding and applying panoramic information. However, research on 360° ISOD has not been widely explored due to the lack of large, complex, high-resolution, and well-labeled datasets. Towards this end, we construct a large scale 360° ISOD dataset with object-level pixel-wise annotation on equirectangular projection (ERP), which contains rich panoramic scenes with not less than 2K resolution and is the largest dataset for 360° ISOD by far to our best knowledge. By observing the data, we find current methods face three significant challenges in panoramic scenarios: diverse distortion degrees, discontinuous edge effects and changeable object scales. Inspired by humans' observing process, we propose a view-aware salient object detection method based on a Sample Adaptive View Transformer (SAVT) module with two sub-modules to mitigate these issues. Specifically, the sub-module View Transformer (VT) contains three transform branches based on different kinds of transformations to learn various features under different views and heighten the model's feature toleration of distortion, edge effects and object scales. Moreover, the sub-module Sample Adaptive Fusion (SAF) is to adjust the weights of different transform branches based on various sample features and make transformed enhanced features fuse more appropriately. The benchmark results of 20 state-of-the-art ISOD methods reveal the constructed dataset is very challenging. Moreover, exhaustive experiments verify the proposed approach is practical and outperforms the state-of-the-art methods.

*Index Terms*—Salient object detection, panoramic dataset, view transformer, distortion.

## I. INTRODUCTION

OMNIDIRECTIONAL images can sample the entire viewing sphere surrounding its optical center, a 360° × 180° FoV [1], [2] and the resolution of an omnidirectional image (ODI) is always several times that of the traditional image, making storing, transmitting and understanding more difficult [3]. Therefore, salient object detection, automatically processing regions of interest and selectively ignoring parts of uninterest, is significant for compressing, transmitting and analyzing 360° panoramic images [4]–[6].

Currently, panoramic datasets are evolving to meet increasing demands in benchmarking and developing 360°-based ISOD models. Li et al. [7] construct the first 360° ISOD dataset 360-SOD with pixel-wise object-level annotation

J. Wu, T. Yu and J. Li are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China. J. Li is also with Peng Cheng Laboratory, Shenzhen, 518000, China.

C. Xia is with Peng Cheng Laboratory, Shenzhen, 518000, China.

Correspondence should be addressed to C. Xia (e-mail: xiachq@pcl.ac.cn) and Jia Li (e-mail: jiali@buaa.edu.cn). Website: http://cvteam.net



Fig. 1. Representative examples of 360° omnidirectional images in ERP form. These three columns show the cases of diverse distortion degrees, discontinuous edge effects and changeable object scales, respectively.

containing 500 equirectangular projection (ERP) images in 512 × 1024 resolution from existing human fixation datasets. Ma et al. [8] collect 1105 ERP images in 546×1024 resolution with object-level annotation. Zhang et al. [2] provide a dataset with object-level and instance-level annotation containing 107 ERP images in 512 × 1024 resolution. However, currently available datasets are relatively small in scale and resolution and less complex in scenarios, which is not enough for further studies and is a primary cause for limited related research. Besides, insufficient training data easily leads to model overfitting. Therefore, it is urgent to break the data bottleneck.

To this end, we construct a new large scale 360° omnidirectional image-based salient object detection (SOD) dataset referred to as ODI-SOD with object-level pixel-wise annotation on ERP to assist studies about 360° ISOD task. The proposed dataset contains 6,263 ERP images with not less than 2K resolution selected from 8,896 panoramic images and 998 videos. The chosen images have the number of salient regions ranging from one to more than ten, the area ratios of salient regions from less than 0.02% to more than 65% and the resolutions from 2K to 8K. More than half of the scenarios are complex and contain diverse objects.

Moreover, through the observation of the dataset, we find that the poor performances of existing state-of-the-art methods [7], [9], [10] can attribute to three prominent challenges, i.e., diverse distortion degrees, discontinuous edge effects and changeable object scales. Distortion varying with the projection position (e.g., the first col in Fig.1) leads to uniform filter

sampling and feature learning difficulties. Edge discontinuity (e.g., the animal is split into two parts on the borders in the second col of Fig.1) makes it difficult to segment complete salient objects on projection borders. Changeable scale objects, especially small/large ones in wide FoV panoramas (e.g., the last cols in Fig.1), make detection and segmentation more difficult.

In fact, to mitigate above challenges in 360° ISOD, researchers have made some related attempts. For example, Li et al. [7] propose a distortion-adaptive module that cuts each ERP image into four image blocks to learn different kernels and design a multi-scale module to integrate context features. Ma et al. [8] put forward a multi-stage ISOD method to handle ERP distortions by using less distorted perspective images and object-level semantical saliency ranking. However, these methods mainly focus on alleviating distortion, ignoring that an advantageous characteristic of a panorama is the continuous complete panoramic field of view. Cutting an ERP image into blocks makes a panorama lose its full panoramic view and may only segment out partial object regions due to complete objects being broken. Perspective images also contain a limited field of view and may be affected by the previous stage's mistakes. Moreover, multi-stage learning and perspective images demanding heavy memory are not friendly to model training/testing.

The FoV of a panorama is $360° \times 180°$ while the binocular visual field of human beings is about $120°$ [11]. To better understand the panorama, humans usually change the viewpoint (e.g., look up/down or left/right) or adjust view distance (e.g, zoom in/out) to obtain more scene information from different perspectives. The whole observing process is smooth and keeps the complete panoramic view. Inspired by the observing behaviors of human beings, in this paper we put forward a solid Sample Adaptive View Transformer (SAVT) module based on various geometric transformations. SAVT contains two sub-modules, View Transformer (VT) and Sample Adaptive Fusion (SAF). Simulating humans' observing process, VT makes different feature transformations based on different ERP center viewpoints or view distances to learn various features under different views. Following VT, SAF generates adaptive weights for different transform branches based on sample characteristics and makes the features fuse better. Combining VT and SAF, the effects of SAVT are threefold: 1) mitigating the effects of discontinuous edges by changing center viewpoints, 2) better locating and segmenting objects in changeable scales by converting view distances to obtain scalable scene information, and 3) heightening the feature toleration of distortion by increasing the distortion diversity. It is different from the methods of adapting distortion or reducing distortion by adjusting filter sampling methods [12], [13]. Benchmark results on 20 state-of-the-art ISOD methods present the proposed dataset is challenging. Moreover, qualitative and quantitative experiments verify the proposed method is effective and outperforms the state-of-the-art methods.

Our contributions are as follows:

- **We construct a new large-scale challenging 360° ISOD dataset named ODI-SOD. It contains 6263 high-resolution ERP images with object-level pixel-wise annotation and is the largest 360° ISOD dataset to the best of our knowledge.**
- **To our best knowledge, we should be the first one to transfer humans' observing process for panoramas to deep feature learning for ERP images.**
- **Inspired by humans' observing behaviors, we propose a novel Sample Adaptive View Transformer (SAVT) module, which keeps the complete panoramic view and mitigates the effects of distortion, edge discontinuity, and changeable scale objects in panoramic scenarios.**
- **We make a benchmark on the proposed dataset using 2D ISOD methods, 360° panoramic ISOD methods and our methods. Our approach outperforms existing state-of-the-art methods.**

## II. RELATED WORK

In this section, we briefly review existing mainstream 360° panoramic datasets and 360° panoramic models.

### A. 360° Panoramic Datasets

Datasets play an important role in object detection tasks such as salient object detection [14], co-salient object detection [15], RGB-D salient object detection [16], [17] and camouflaged objects detection [18], [19]. For example, in the 2D domain, the remarkable progress of the ISOD task benefits much from the construction of representative datasets [14], [20]–[29]. Early datasets are often limited in the number of images or scene complexity [14], [20]–[23]. Whereafter, two large-scale and challenging datasets XPIE [24] and DUTS [25] are introduced to overcome preceding shortcomings. Besides, salient object datasets with instance-level annotation are proposed [26]–[29] to promote the research.

Recently, some researchers [2], [7], [8], [30]–[42] turn attention to saliency studies in 360° panoramic scenarios. While most datasets only provide either eye-fixation groundtruth data for saliency prediction or bounding box groundtruth for object detection, which can promote salient object detection but is not enough for accurate pixel-wise salient object segmentation in panoramic scenarios. Therefore, three small-scale omnidirectional image-based SOD datasets with pixel-wise annotation, i.e., 360-SOD [7], F-360iSOD [2] and 360-SSOD [8], are successively proposed for 360° ISOD. 360-SOD [7] is the first 360° SOD dataset and has 500 ERP images with pixel-wise object-level annotation based on human fixation groundtruth. F-360iSOD [2] is the first 360° SOD dataset providing pixel-wise object level and instance level binary masks and contains 107 ERP images with 1,165 salient objects. The latest dataset 360-SSOD [8] has 1,105 semantically balanced ERP images with only object-level masks. To our best knowledge, they are the datasets available for the 360° ISOD task.

However, the available datasets are insufficient in number or the scene complexity to understand the real-world panoramic scenarios. It is expected that a large-scale high-resolution dataset with rich and complex scenarios is built to alleviate data constraints. Rich and complex scenarios are closer to the

real world, the large-scale number is helpful for training models, and the high-resolution represents the detail information better. Thus, in this paper, we introduce a large-scale 360° ISOD dataset with high-resolution and complex scenarios. The general information of representative datasets for 2D ISOD and 360° panoramic ISOD is shown in Tab.I.

### B. 360° Panoramic Models

Large-scale 2D ISOD datasets such as DUTS [25] and XPIE [24] have extensively promoted the development of CNN-based ISOD methods [43]–[51]. However, 360° panoramic SOD models are minimal due to insufficient object-level pixel-wise annotation [7], [8], [52]. In [7], a distortion adaptive module for 360° ISOD is proposed to alleviate the distortion effects from the equirectangular projection by cutting the input equirectangular image into several blocks to deal with different regions with various parameters. [8] put forward a multi-stage coarse-to-fine SOD method for ODIs to handle the effects of distortions and complex scenarios using perspective images with less distortion and object-level semantical saliency ranking. Moreover, [52] uses ERP images and much less distorted cube-map images as network input to extract and fuse features adaptively. Yet, they only focus on alleviating distortion, ignoring the effects of edge discontinuity and panoramic FoV. Similar problems also exist in other 360°-based tasks such as saliency prediction [53], object detection [36], panoramic semantic segmentation [54], 3D room layout [12], and dense prediction [13].

For 360° scenarios, the continuous panoramic view is an advantageous characteristic. Cutting a panorama into blocks or using perspective images can lose the original panoramic view and may bring more discontinuous edges, especially for changeable scale objects in complex panoramic scenes. Therefore, in this paper, we propose a 360° ISOD model with the consideration of distortion, edge discontinuity and changeable scale objects in panoramic FoV.

## III. DATASET

There are two limitations in the existing three panoramic datasets. Firstly, the most extensive dataset only contains 1105 images, which is insufficient to train a general deep network and easily leads to overfitting. Secondly, the image resolutions of the datasets are not satisfying for further research on complex 360° scenarios. In this section, a new large-scale dataset named ODI-SOD[1] is introduced from the aspects of dataset construction, dataset statistics and analysis.

### A. Dataset Construction

1) Dataset Collection: The dataset ODI-SOD comprises 1,151 images collected from the Flickr website and 5,112 video frames selected from YouTube. All panoramas are in equirectangular projection format (the ratio of height and width is strictly 2:1), and the resolutions are not less than 2K. During collection, we search panoramic sources on Flickr and

[1]The ODI-SOD dataset will be published and can be downloaded via https://github.com/iCVTEAM/ODI-SOD.git

YouTube with different object category keywords (e.g., human, dog, building) referring to MS-COCO classes [55] to cover various real-world scenes. In this way, we collect 8,896 images and 998 videos, including different scenes (e.g., indoor, outdoor), different occasions (e.g., travel, sports), different motion patterns(e.g., moving, static), and different perspectives. Then, all videos are sampled into keyframes, and the unsatisfactory images or frames (e.g., without salient objects, low quality) are dropped out. Finally, 6,263 ERP source samples are selected for the subsequent annotation.

2) Salient Object Annotation: Considering most 360° scenarios are complex and contain more than one object, there always exist some ambiguous objects between saliency and not saliency. It is necessary to select salient objects before time-consuming annotation. Firstly, we require five researchers to judge object saliency and select salient objects by voting. Secondly, annotation aspects manually label binary masks based on the chosen salient objects. Finally, five researchers cross-check the binary masks to ensure accurate pixel-wise object-level annotations. Some sample pairs have been shown in Fig.1

3) Dataset Split: The dataset is divided into a test set with 2,000 images and a train set with 4,263 images for deep network training. Note that all source frames from the same video are divided into the same set, train set or test set, and other source images/frames are randomly divided.

### B. Dataset Statistics

To explore the main characteristics of the proposed dataset and compare it with existing 360° ISOD datasets, we make statistics on typical attributes of salient regions, including edge discontinuity, distortion degree and max FoV coverage.

1) Discontinuity of Salient Object Regions: Different from 2D images, the left boundary and right boundary of 360-degree images in ERP format are connected [35]. For a panorama, if the central meridian crosses target salient object regions, the complete and continuous salient object regions will be divided into two discontinuous parts by the left and right boundaries of its ERP image. Here, the discontinuity of salient object regions at the boundaries is called discontinuous edge effects, which is also one of the major challenges. For the ERP images with discontinuous edge effects, it is usually more difficult to obtain complete segmentations due to the forced separation in space. Thus, it is significant to make statistics about image proportions with discontinuous edge effects. Fig.2 presents the percentage of images with edge discontinuity and without edge discontinuity for the existing 360° ISOD datasets. It can be seen that our dataset has a more balanced distribution and a larger number of images with edge discontinuity compared with other datasets, which is beneficial for exploring discontinuous edge effects.

2) Distortion of Salient Object Regions: The distortion degrees of salient object regions usually change with their locations, reaching a maximum at the polar regions and a minimum on the equator. Given an ERP image $I$ and its binary groundtruth $G$ with width $w$ and height $h$, the coordinate of point $P$ can be represented as $P = (x, y)$ on the 2D

TABLE I
REPRESENTATIVE DATASETS FOR ISOD.

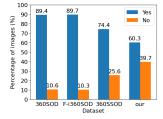| Task | Dataset | Year | #Image | #GT | Res.[min, max] | GT Level | Description |
|---|---|---|---|---|---|---|---|
| 2D ISOD | ECSSD [14] | CVPR'13 | 1,000 | 1,000 | [139, 400] | obj. | includes many semantically meaningful but structurally complex images |
| | DUT-OMRON [20] | CVPR'13 | 5,168 | 5,168 | [139, 401] | obj. | one or more salient objects and relatively complex background |
| | PASCAL-S [21] | CVPR'14 | 850 | 850 | [139, 500] | obj. | multiple salient objects, total 1296 relatively salient object instances |
| | MSRA10K [22] | TPAMI'14 | 10,000 | 10,000 | [165, 400] | obj. | most are with only one salient object and simple background. |
| | HKU-IS [23] | CVPR'15 | 4,447 | 4,447 | [100, 500] | obj. | most have either low contrast, complex background or multiple salient objects |
| | XPIE [24] | CVPR'17 | 10,000 | 10,000 | [128, 300] | obj. | covers many complex scenes with different numbers, sizes and positions of salient objects |
| | DUTS [25] | CVPR'17 | 15,572 | 15,572 | [100, 500] | obj. | from the ImageNet DET set and the SUN data set, very challenging scenarios |
| | ILSO-1K [26] | CVPR'17 | 1,000 | 1,000 | [142, 400] | obj.& ins. | contains instance-level salient objects annotation but has boundaries roughly labeled |
| | SOC [27] | ECCV'18 | 6,000 | 6,000 | [161, 849] | obj.& ins. | with salient and non-salient objects from more than 80 common categories |
| | SIP [28] | TNNLS'20 | 929 | 929 | [744, 992] | obj.& ins. | salient person samples that cover diverse real-world scenes |
| | ILSO-2K [29] | CVIU'21 | 2000 | 2000 | [142,400] | obj.& ins. | most contain multiple salient object instances, complex background, or low color contrast. |
| 360° ISOD | 360-SOD [7] | JSTSP'19 | 500 | 500 | [409, 1024] | obj. | ERP images from five panoramic video datasets with fixation groundtruth |
| | 360-SSOD [8] | TVCG'20 | 1105 | 1105 | [546, 1024] | obj. | ten categories, ERP images from 677 panoramic videos |
| | F-360iSOD [2] | ICIP'20 | 107 | 107 | [1024, 2048] | obj.& ins. | 107 panoramic images, 1,165 salient objects, 9 images without any salient object annotations |
| | ODI-SOD | 2022 | 6263 | 6263 | [1024, 11264] | obj. | 6263 panoramic images captured in real-world scenes and each image has pixel-wise annotation |



Fig. 2. The percentage statistics of images with edge discontinuity (marked for Yes) and without edge discontinuity (marked for No).
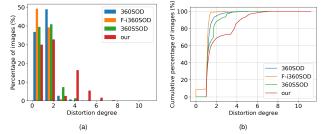


Fig. 3. The statistical distribution of images with different distortion degrees. (a) The histogram distribution. (b) The cumulative distribution.

pixel plane or $P = (\lambda, \varphi)$ on the sphere surface using longitude and latitude, in which $x \in X = \{0, 1, ..., w - 1\}, y \in Y = \{0, 1, ..., h - 1\}, \lambda \in \Lambda = [-180°, 180°], \varphi \in \Phi = [-90°, 90°], \varphi = E(y)$ and $E$ is the inverse projection operator. Based on $G$, we can get the salient area $a_y$ of each row on the ERP pixel plane and the corresponding area $s_\varphi$ on the sphere surface obtained, in which $s_\varphi = a_y \cdot \cos \varphi$. To quantize the distortion, for each image, we define the distortion degree $D$ of the salient object regions as follows:

$$D = \frac{1}{n} \sum_{j=0}^{n-1} \frac{a_{y_j}}{s_{\varphi_j}} = \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{\cos \varphi_j}, \quad (1)$$

here, $\varphi_j = E(y_j)$ and $(y_j, \varphi_j) \in Q = \{(y_j, \varphi_j)|y_j \in Y, \varphi_j \in \Phi, s_{\varphi_j} > 0\}$, and $n$ is the number of elements in set $Q$. From Eq.1, we can find the distortion degree of salient object regions in an ERP image mainly depends on the vertical

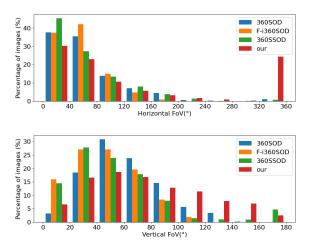Fig. 4. Examples of image and groundtruth pairs with different distortion degrees.



Fig. 5. Horizontal and vertical FoV of salient object regions.

FoV coverage and horizontal-wise area ratio of salient regions, which is a general measure of the distortion degree and has no direct relation with the number and area of salient regions. Resizing all images in datasets to the same resolution and calculating their distortion values, the statistical distribution of images with different distortion degrees are counted and shown in Fig.3. We can find that our dataset has a larger range of distortion degree distribution than other datasets. For example, in Fig.3(b) the distortion degrees of our dataset range from 1 to about 11, while the best of others range from 0 to about 7. Besides, the proportion of images with large distortion degrees in our dataset is much larger than that in other datasets. For example, in Fig.3(a) our dataset still has obvious distribution when the distortion degree is larger than 3. In Fig.3(b), for other datasets the percentages of images with distortion degrees less than 3 reach more than 90%, which means the percentages of images with distortion degrees larger than 3 are less than 10%, while for our dataset, the percentage of images with distortion degrees larger than 3 is about 30%. Moreover, in Fig.4 we provide some sample pairs with different distortion degrees, presenting the reasonability of the above distortion degree formulation.

*3) FoV Coverage of Salient Object Regions:* ODIs can sample the entire viewing sphere surrounding its optical center,
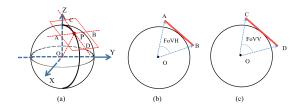


Fig. 6. A diagram of the viewport. P is the tangent point of the sphere and viewport, i.e., the viewpoint. Points A, B, C, and D are the center points of viewport edges. FoVH and FoVV are the horizontal and vertical fields of view, respectively.

a $360° \times 180°$ FoV [1], [2]. Each salient object region covers a horizontal FoV and a vertical FoV, and the covered horizontal/vertical FoV can reflect the horizontal/vertical scale of the salient object region. Usually, the vertical FoV coverages have more balanced distributions than the horizontal FoV coverages since salient regions are stretched more horizontally. To obtain holistic distributions of each dataset, calculate the max horizontal/vertical FoV coverage of salient regions in each ERP image and plot the histogram distribution in Fig.5. We find that the percentage of images decreases with the covered horizontal FoV increasing, and there are fewer images when the covered horizontal FoV is larger than $240°$ except in our dataset. For vertical FoV, the percentage of images reaches the maximum in $[20°, 60°]$. The max vertical FoV coverages of most images in other datasets are smaller than $120°$. Compared with other datasets, our dataset has more balanced and smooth distributions. It has a larger percentage of images with targets covering large FoVs, which indicates our dataset is more challenging due to the general existence of salient regions with different scales.

### C. Dataset Analysis

From the dataset statistics, we find that discontinuous edge effects and different degrees of distortions are unavoidable due to the equirectangular projection and that different scales of salient regions are very common in complex panoramic scenes. In some cases, these characteristics can occur at the same time, which makes discontinuous edge effects, diverse distortion degrees and changeable object scales become main challenges of the $360°$ ISOD task. Therefore, it is necessary to design an effective model to solve above problems.

### IV. APPROACH

To overcome above problems, we present our overall approach as shown in Fig.7, which consists of the encoder, decoder and the proposed Sample Adaptive View Transformer (SAVT) module that has two sub-modules View Transformer (VT) and Sample Adaptive Fusion (SAF). To better understand the two modules, we first introduce basic concepts in Sec. IV-A. Then, we describe the overall framework of our method in Sec. IV-B. Subsequently, Sec. IV-C illustrate the proposed SAVT in detail.

### A. Preliminary

In this part, we briefly introduce the process and characteristics of the equirectangular projection and explain the viewport and viewpoint used in panoramas and Möbius transformations.
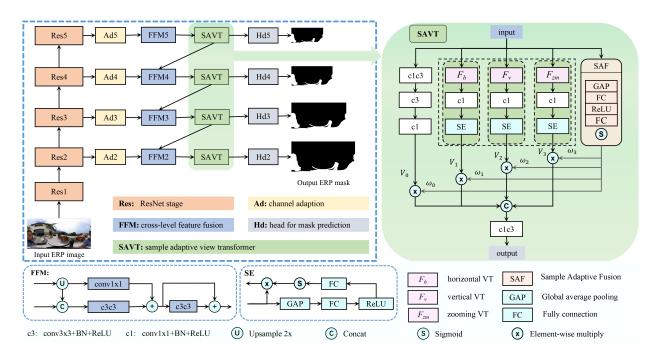
Fig. 7. The framework of our method and the proposed module SAVT with two sub-modules, View Transformer (VT) and Sample Adaptive Fusion (SAF). Specifically, VT contains three sub-branches $F_h$, $F_v$ and $F_{zm}$. $F_h$ means horizontal view transformer, $F_v$ means vertical view transformer and $F_{zm}$ means zooming view transformer.

*1) Viewport:* As shown in Fig.6, in panoramic vision, when looking at the point $P(\lambda, \varphi)$ from the sphere center $O$ with the horizontal and vertical field of view FoVH and FoVV, respectively, we can see a region $R$ of the sphere surface, and $P$ is the center of view (i.e., viewpoint). The points on the sphere region $R$ can be projected to a rectangular plane tangent to the sphere surface at point $P$ by gnomonic projection. The tangent plane is defined as the viewport as [56], [57] do. The distance from $O$ to $P$ is called view distance in the study.

*2) Möbius Transformation:* Möbius transformations are one-to-one, onto and conformal (angle preserving) maps of the so-called extended complex plane [58]. The extended complex plane is given by $\mathbb{C}_\infty = \mathbb{C} \cup \infty$. A Möbius transformation $f : \mathbb{C}_\infty \to \mathbb{C}_\infty$ is a map

$$f(z) = \frac{az+b}{cz+d}, \quad a,b,c,d \in \mathbb{C}, ad-bc \neq 0, \quad (2)$$

where a,b,c and d are constant complex numbers satisfying $ad - bc \neq 0$.

The Riemann sphere is a model of $\mathbb{C}_\infty$, which is homeomorphic to the two-dimensional sphere [59]

$$S^2 = \{(x_s, y_s, z_s) \in \mathbb{R}^3 | x_s^2 + y_s^2 + z_s^2 = 1\}. \quad (3)$$

We identify the complex plane $\mathbb{C}$ with the equitorial plane $x_3 = 0$, and set the North pole $N = (0, 0, 1)$. If the line from $N$ to $P$ intersects the complex plane in exactly one point $z \in \mathbb{C}$, then the map $SP : S^2 \setminus N \to \mathbb{C}$ which assigns a point $P \in S^2$ to the point $z \in \mathbb{C}$ is called the stereographic projection [58], which is a bridge between $R^3$ and $\mathbb{C}_\infty$. For $P = (x_s, y_s, z_s) \in S^2 \setminus N$ and $z = x + iy \in \mathbb{C}$, $SP$ is given by

$$SP : (x_s, y_s, z_s) \to \frac{x_s}{1 - z_s} + i\frac{y_s}{1 - z_s}. \quad (4)$$

For ODIs, ERP is just one of the projection formats. The vanilla representation is the sphere surface representation, also called visible sphere, fully representing the original field of view with $360°$ longitude by $180°$ latitude. Therefore, Möbius transformations can be applied to ODIs, which is vital for the proposed method.

*B. Overall Framework*

To take advantage of existing mature 2D CNNs, we take into account the classic U-shape structure and add the proposed tailored delicate SAVT module aiming at ODIs. The overall framework is shown in Fig. 7. The input and output are both in the ERP format.

For the encoder, the backbone network uses ResNet-50 [60] removed the last global pooling and fully connected layers for the pixel-level prediction. For the decoder, the output features of the encoder pass through the channel adaption modules and feature fusion modules FFM, and then transmit to SAVT. Each FFM connecting with $\{Res2, Res3, Res4\}$ fuses the features of the current stage and adjacent higher stage into enhanced features for SAVT in the current stage. The FFM in stage 5 ignores the upsample interpolation and concat operations at the entrance. Each SAVT connects with a mask head consisting of a convolution layer with kernel $3 * 3$ as the channel compression layer and an upsampling interpolation operation. All mask heads are used for side-output supervision in the training stage. The progressive strategy from coarse to fine is beneficial for SOD tasks.

The proposed SAVT contains two parallel sub-modules, View Transformer (VT) and Sample Adaptive Fusion (SAF). VT has three branches $F_h, F_v, F_{zm}$ corresponding to different transformations to simulate the human observing process of

changing viewpoints or view distances, and SAF is used to adjust the output values of other parallel branches. Next, we present it in detail.

### C. Sample Adaptive View Transformer

*1) View Transformer Types:* No matter the ODIs are displayed in the desktop setting or head-mounted VR, what can be seen is very limited at each moment. We have to change our viewpoint or adjust view distance to obtain more information. Inspired by this, for ERP image processing, we introduce rotation and zooming, the two kinds of transformation, to simulate the observing process of looking left and right (branch $F_h$), up and down (branch $F_v$), far and near (branch $F_{zm}$).

*2) View Transformer Formulation:* An ODI represented as the Riemann sphere can use different Möbius transformations, making the panoramic scene keep continuous in panoramic view after transformation.

About rotation, a map $f : \mathbb{C}_\infty \to \mathbb{C}_\infty$ is called a rotation of $\mathbb{C}_\infty$ if the map $SP^{-1} \circ f \circ SP : S^2 \to S^2$ is a rotation [58]. Möbius transformations represent rotations if and only if $c = -\bar{b}, d = \bar{a}$, and $ad - bc = |a|^2 + |b|^2 = 1$, i.e.,

$$f(z) = \frac{az + b}{-\bar{b}z + \bar{a}}, \quad a, b, c, d \in \mathbb{C}, ad - bc = 1. \quad (5)$$

For convenience, under a rotation of an angle $\theta \in [0, 2\pi]$ about the axis passing through the origin in the direction along the vector $\mathbf{L} = (l, m, n)$, based on the formulas of stereographic projection and Riemann sphere [61], the complex number $a, b$ in $f(z)$ can be derived as follows:

$$a = cos(\theta/2) + i \cdot n \cdot sin(\theta/2), \quad (6)$$
$$b = (m - i \cdot l)sin(\theta/2). \quad (7)$$

If $\mathbf{L} = (0, 0, 1)$, then $a = cos(\theta/2) + i \cdot sin(\theta/2), b = 0$, then there is

$$f(z) = \frac{az}{\bar{a}} = e^{i\theta}z, \quad (8)$$

namely the canonic *elliptic* Möbius transformation, which can simulate looking left or right. Similarly, if $\mathbf{L} = (0, 1, 0)$, then it can simulate looking up or down.

About zooming, we set $c = 0$ and simplify Eq.2 as follows:

$$f(z) = az, \quad a \in \mathbb{C}. \quad (9)$$

To facilitate, write $a = \rho e^{i\theta}$. When $\theta = 0, \rho < 1, f(z)$ is an origin-centered contraction, and when $\theta = 0, \rho > 1, f(z)$ is an origin-centered expansion. The Möbius transformations with contraction or expansion are called *hyperbolic*. We can change the zooming center by rotating the target center to the origin and inversely rotating it back after zooming.

Before applying to the feature space, we take the transform process on an ERP image as an example for a simplified description and better visualization of the geometric transformation. For each point $P_e(u, v)$ on the ERP pixel plane, there is a corresponding point $P_s(\lambda, \varphi)$ ($\lambda$ is longtitude, and $\varphi$ is latitude) on the sphere surface. The coordinate of $P_s$ in $\mathbb{R}^3$ is $P(x_s, y_s, z_s)$. Usually, the $P_s(0, 0)$ is projected to the center
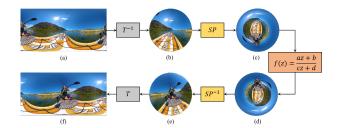


Fig. 8. Transform steps in View Transformer.



Fig. 9. Some transformation examples by View Transformer. Each row from top to bottom corresponds to the original images, horizontal transformation, vertical transformation, and zooming transformation. From left to right, the horizontal transformation parameters are $\theta = 150°, 100°$ and $180°$; the vertical transformation parameters are all $\theta = 30°$; the zooming transformation parameters are $\rho = 1.5(O = (0, 1, 0)), \rho = 0.5(O = (0, 1, 0))$ and $\rho = 1.5(O = (-1, 0, 0))$.

of equirectangular image. The transform relationship can be represented as

$$P_e(u, v) = T(P(x, y, z)), \quad (10)$$

$$P(x, y, z) = T^{-1}(P_e(u, v)), \quad (11)$$

where $T(\cdot)$ is the transform function from the sphere surface to the ERP pixel plane, and $T^{-1}(\cdot)$ is its inverse transform function.

As shown in Fig.8, for a given ERP image, the whole process is divided into five steps: 1) first back-project the ERP pixel plane to the sphere surface; 2) through stereographic projection $SP$, get its representation in the extended complex plane $\mathbb{C}_\infty$, 3) make Möbius transformations in $\mathbb{C}_\infty$; 4) back-project to the Riemann sphere after transformations; 5) project to the ERP pixel plane from the sphere surface. The simplified formula is as follows:

$$P'_e = T(SP^{-1}(f(SP(T^{-1}(P_e))))). \\ = F(P_e). \quad (12)$$

Here, $P'_e$ is the point on the ERP image after view transformation. The whole process of transformation is reversible, represented as $F^{-1}(\cdot)$. By doing this, for an ERP image, we can get the transformation images under different views, and the panoramic views are kept simultaneously. Fig. 9 shows some transformation examples of images. It can be seen that transformations keep the complete panoramic view and obtain appearances under diverse views.
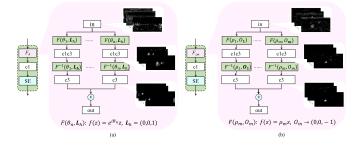
Fig. 10. View Transformer branches. (a) Horizontal rotation branch with $F_h$. (b) Zooming branch with $F_{zm}$. In (a), $\theta_i, i \in [1, n]$ is the rotation degree and $L_h$ is the rotation axis. In (b), $\rho_i, i \in [1, m]$ is the zooming factor and $O_m$ is zooming center.

*3) Refine the Feature Space:* In the study, we use the transforming process to refine the feature space. Then, each $P_e$ or $P_e'$ corresponds to a feature position before or after transformation. The feature value $v_p$ at the feature point $p = (i, j, k)$ is mapped to the target position $p' = (i', j', k)$ through transformation, in which $i, i' \in [0, h-1], j, j' \in [0, w-1]$, $k \in [0, c-1]$, $h$ is the height of the input feature, $w$ is the width and $c$ is the channels. Theoretically, in the pure geometric transformation, the feature value is unchanged. The geometric transformations contribute to rich feature appearances under different views, which is view-aware. Thus, by feat of the transformed features, we can learn more appropriate features and fuse them after inverse transformation.

Fig.10 shows more details about branches of VT, in which the vertical branch is not contained because it is similar to the horizontal branch except for the vector $L_v = (0, 1, 0)$. $F_h, F_v, F_{zm}$ contain different numbers of sub-branches, and each sub-branch corresponds to different transform parameters. For $F_h$, $L_h$ is fixed and the degree $\theta_i, i \in [1, n]$ is the horizontal rotation degree of looking left or right. For $F_v$, $L_v$ is fixed and the degree parameter is the vertical rotation degree of looking up or down. For $F_{zm}$, $O_m$ is like our viewpoint and $\rho_i, i \in [1, m]$ controls looking near or far.

*4) Sample Adaptive Fusion:* To make better use of these transformation features, we perform an adaptive fusion of these features to adapt to different samples. This fusion process is expected to be simple and efficient. Here, we use a SENet block [62] to realize it by learning an adaptive weight for each type of transformation branch and the original learning branch (see Fig 7). Then fuse the weighted features by a concat operation in the channel dimension, as follows:

$$V_f = Concat(\omega_k \cdot V_k), \quad k = 0, 1, 2, 3, \tag{13}$$

where $V_0$ corresponds to the original feature learning branch without any geometric transformation, $V_1$ and $V_2$ correspond to the two rotation branches, $V_3$ corresponds to the zooming branch, and $\omega_k$ is the function of the original input feature which depends on the input sample. Thus, we called the process Sample Adaptive Fusion. Through SAF, the transformed features can be adaptively fused and better represent the current sample.

Fig.11 shows the gradient class activation maps (CAMs) [63] of some representative samples with objects

that are obviously distorted or with discontinuous edge effects or on large or small scales. From Fig.11 we find that the output features (the forth row) of SAVT are much better than its input (the third row), which indicates SAVT is effective. Specifically, there always exists at least one transform branch outputting better features in some regions for different samples, which suggests that VT offers diverse and helpful candidate features. Moreover, the output features of different branches are integrated better by SAF adjusting weights, which shows VT and SAF are combined effectively.
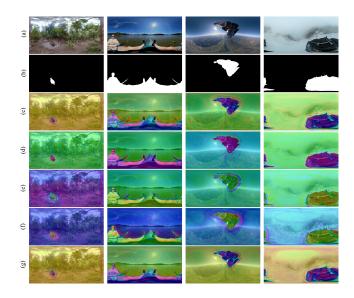


Fig. 11. Visual examples of gradient CAMs of the SAVT module. (a) Input images. (b) Groundtruth maps. (c) CAMs based on input features of SAVT. (d) CAMs based on output features of SAVT. (e) CAMs based on output features of the horizontal rotation branch. (f) CAMs based on output features of the vertical rotation branch. (g) CAMs based on output features of the zooming branch.

## V. EXPERIMENTS

In the section, we first benchmark current state-of-the-art 2D-based SOD methods and $360°$-based SOD methods on the proposed dataset ODI-SOD. Then we choose representative methods to train on our dataset and compare with the proposed method. Furthermore, we verify and analyze the effectiveness of the proposed method in the ablation study.

### A. Experimental Settings

*1) Dataset and Evaluation Metrics:* We conduct experiments on the proposed dataset ODI-SOD, which contains 4,263 training images and 2,000 testing images. To quantitatively evaluate the performance of methods, we utilize the common metrics for SOD, namely the mean absolute error (MAE), F-meansure ($F_\beta$), weighted F-measure($wF_\beta$) [64], max F-measure ($maxF$), S-measure ($S_m$) [65], E-measure ($E_m$) [66], [67]. F-measure indicates the trade-off result between precision and recall and here we set $\beta^2 = 0.3$ to emphasize more precision than recall as in [68].

TABLE II
BENCHMARKING RESULTS OF THE SOTA METHODS ON THE ODI-SOD TEST SET BEFORE TRAINING BY ODI-SOD TRAIN SET.

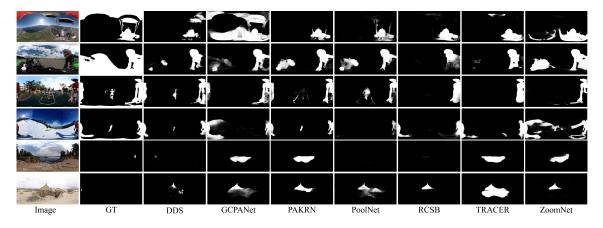| Methods | Year | Type | Backbone | Train set | MAE↓ | $F_\beta$↑ | $wF_\beta$↑ | $S_m$↑ | $E_m$↑ | $maxF$↑ |
|---------|------|------|----------|-----------|------|-----------|-------------|--------|--------|---------|
| GCPANet [46] | 2020 AAAI | 2D SOD | ResNet 50 | DUTS-TR | 0.128 | 0.460 | 0.428 | 0.647 | 0.669 | 0.568 |
| MINet-R [45] | 2020 CVPR | 2D SOD | ResNet 50 | DUTS-TR | 0.123 | 0.435 | 0.399 | 0.624 | 0.664 | 0.528 |
| ITSD [69] | 2020 CVPR | 2D SOD | ResNet 50 | DUTS-TR | 0.137 | 0.450 | 0.427 | 0.635 | 0.655 | 0.538 |
| F3Net [70] | 2020 AAAI | 2D SOD | ResNet 50 | DUTS-TR | 0.133 | 0.423 | 0.387 | 0.615 | 0.655 | 0.519 |
| DFI [71] | 2020 TIP | 2D SOD | ResNet 50 | DUTS-TR | 0.108 | 0.460 | 0.430 | 0.654 | 0.674 | 0.570 |
| PFSNet [72] | 2021 AAAI | 2D SOD | ResNet 50 | DUTS-TR | 0.141 | 0.421 | 0.388 | 0.609 | 0.649 | 0.514 |
| CTDNet [51] | 2021 MM | 2D SOD | ResNet 50 | DUTS-TR | 0.138 | 0.423 | 0.389 | 0.610 | 0.662 | 0.509 |
| VST [50] | 2021 ICCV | 2D SOD | T2T-ViT | DUTS-TR | 0.135 | 0.428 | 0.402 | 0.621 | 0.656 | 0.518 |
| PAKRN [9] | 2021 AAAI | 2D SOD | ResNet 50 | DUTS-TR | 0.092 | 0.556 | 0.518 | 0.694 | 0.729 | 0.642 |
| DCN [73] | 2021 TIP | 2D SOD | ResNet 50 | DUTS-TR | 0.125 | 0.417 | 0.383 | 0.613 | 0.648 | 0.514 |
| SOD100K [74] | 2021 TPAMI | 2D SOD | ResNet 50 | DUTS-TR | 0.198 | 0.288 | 0.245 | 0.544 | 0.543 | 0.401 |
| PSGLoss [75] | 2021 TIP | 2D SOD | ResNet 50 | DUTS-TR | 0.116 | 0.439 | 0.392 | 0.616 | 0.675 | 0.521 |
| SCASOD [76] | 2021 ICCV | 2D SOD | ResNet 50 | DUTS-TR | 0.083 | 0.455 | 0.391 | 0.625 | 0.582 | 0.475 |
| FastSaliency [77] | 2021 TIP | 2D SOD | ResNet 50 | DUTS-TR | 0.185 | 0.319 | 0.287 | 0.557 | 0.575 | 0.429 |
| PurNet [78] | 2021 TIP | 2D SOD | ResNet 50 | DUTS-TR | 0.119 | 0.436 | 0.400 | 0.622 | 0.678 | 0.535 |
| PoolNet [79] | 2022 TPAMI | 2D SOD | ResNet 50 | DUTS-TR | 0.101 | 0.466 | 0.419 | 0.647 | 0.687 | 0.576 |
| RCSB [80] | 2022 WACV | 2D SOD | ResNet 50 | DUTS-TR | 0.108 | 0.490 | 0.427 | 0.630 | 0.692 | 0.561 |
| ZoomNet [10] | 2022 CVPR | 2D SOD | ResNet 50 | DUTS-TR | 0.120 | 0.465 | 0.429 | 0.644 | 0.670 | 0.558 |
| TRACER [81] | 2022 AAAI | 2D SOD | ResNet 50 | DUTS-TR | 0.099 | 0.460 | 0.418 | 0.630 | 0.691 | 0.530 |
| DDS [7] | 2019 JSTSP | 360° SOD | ResNet 50 | 360-SOD | 0.070 | 0.553 | 0.493 | 0.694 | 0.751 | 0.648 |



Fig. 12. Visual testing examples of representative state-of-the-art algorithms before finetuning on the ODI-SOD train set.

*2) Implementation Details:* In the training stage, we use the pre-trained ResNet-50 model [60] to initialize the parameters of the feature encoder and use a standard stochastic gradient descent algorithm to train the whole network end-to-end with the cross-entropy loss and IoU loss. In our network encoder, the initial learning rate is set to 0.05 with a weight decay of 0.0005 and momentum of 0.9. For the rest layers, the learning rates are ten times the encoder. We train the proposed method with a mini-batch of size 16 about 64 epochs by a single GTX 3080 GPU. In the testing stage, only the output of $Hd2$ in Fig.7 is used for the final prediction result. In both training and testing, the input images are resized to $512\times256$ resolution for comparison with other 360° SOD methods.

### B. Benchmarking Results

To verify the challenges of the proposed dataset ODI-SOD, in Tab.II we list the performances of 20 state-of-the-art (SOTA) 2D SOD and 360° SOD methods on our test set without finetuning on our train set. The methods include GCPANet [46], MINet-R [45], ITSD [69], F3Net [70], DFI [71], PFSNet [72], CTDNet [51], VST [50], PAKRN [9], DCN [73], SOD100K [74], PSGLoss [75], SCASOD [76], FastSaliency [77], PurNet [78], PoolNet [79], RCSB [80], ZoomNet [10], TRACER [81] and DDS [7]. From Tab.II we find that all listed methods perform not well on the ODI-SOD test set including the 360°-based method DDS [7] trained by the 360-SOD train set, which suggests that currently available models have poor generalization ability over the proposed dataset. It comes down to two reasons. Firstly, a gap exactly exists between 2D SOD datasets and 360° omnidirectional SOD datasets, which makes the outstanding 2D methods have sharp drops in performance. Secondly, the proposed dataset is very challenging and beyond the cognitive capabilities of the existing datasets and models.

For further comprehensive analysis, some testing results of representative superior SOTA algorithms in Tab.II are shown in Fig.12. From the prediction maps, we find that the less distorted or evident salient target regions can be handled by most of the methods. In contrast, the severely distorted regions can easily lead to segmentation failure due to their apparent differences from existing knowledge and perception. The target objects with discontinuous edge effects and the

TABLE III
PERFORMANCE COMPARISON OF OUR METHOD, 360°-BASED METHOD AND THE TOP-5 2D-BASED SOTA METHODS ON THE ODI-SOD TEST SET AFTER
FINETUNING ON THE ODI-SOD TRAIN SET.

| Methods | Year | Type | Params (M) | MACs (G) | MAE↓ | $F_\beta$↑ | $wF_\beta$↑ | $S_m$↑ | $E_m$↑ | $maxF$↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| PAKRN [9] | 2021 AAAI | 2D SOD | 141.06 | 228.85 | 0.106 | 0.408 | 0.410 | 0.632 | 0.611 | 0.727 |
| PoolNet [79] | 2022 TPAMI | 2D SOD | 69.56 | 229.10 | 0.045 | 0.631 | 0.652 | 0.804 | 0.798 | 0.790 |
| RCSB [80] | 2022 WACV | 2D SOD | 27.25 | 454.23 | 0.067 | 0.590 | 0.488 | 0.675 | 0.755 | 0.652 |
| ZoomNet [10] | 2022 CVPR | 2D SOD | 32.38 | 90.44 | 0.039 | 0.712 | 0.689 | 0.805 | 0.863 | 0.804 |
| TRACER [81] | 2022 AAAI | 2D SOD | 3.90 | 2.78 | 0.044 | 0.667 | 0.648 | 0.770 | 0.850 | 0.740 |
| DDS [7] | 2019 JSTSP | 360° SOD | 27.23 | 60.36 | 0.045 | 0.630 | 0.635 | 0.791 | 0.808 | 0.761 |
| our | 2022 | 360° SOD | 56.86 | 42.87 | 0.035 | 0.759 | 0.738 | 0.831 | 0.886 | 0.822 |

small-scale or large-scale objects are also difficult to be completely segmented out. It illustrates that severe distortions, discontinuous edge effects and changeable scales are three major challenges in the proposed 360° dataset.

### C. Comparison with State-of-the-arts

*1) Quantitative Evaluation:* To demonstrate the effectiveness of the proposed method, we selected the Top-5 in performance from 2D SOD methods with available training code in Tab.II, i.e., PAKRN [9], PoolNet [79], RCSB [80], ZoomNet [10] and TRACER [81]. Then, for a fair comparison, we finetune these five models, DDS [7] and our proposed model on the ODI-SOD train set. After finetuning, their evaluation results on the ODI-SOD test set are listed in Tab.III. We can see that the overall metric scores are much better than those before finetuning in Tab.II for the selected methods except for PAKRN [9]. One primary reason may be that PAKRN [9] needs multi-stage joint training, which is not easy for a new task. Surpassing PAKRN [9] and DDS [7], ZoomNet [10] becomes the best one, but the performance is nowhere near as good as on 2D SOD datasets, which illustrates the challenge of the proposed dataset again. Compared with the other methods, our method demonstrates sustained advantages and significant improvements on all the listed metrics and has become the new state-of-the-art. It verifies that the proposed method is effective for the 360° ISOD task.

We sort the ODI-SOD test samples by different attributes, including the target foreground ratio, max horizontal FoV, distortion degree and discontinuous edge effect. To further evaluate the methods' performance variation trends with the attributes, from the finetuned models in Tab.III we choose the best 2D model ZoomNet [10], the 360° model DDS [7] and our model to make predictions, statistics and analysis. Based on the predicted maps we compute the methods' $wF_\beta$ scores on each sample. Then, the statistic scores about discontinuous edge effects are shown in Tab.IV, and the scores about the other attributes are plotted as broken line graphs in Fig. 13. For better visualization and analysis, the lines are smoothed by a moving averaging window with size 50, and the secondary Y-axis is the attribute values.

From Tab.IV we observe that our method outperforms other methods on all listed criteria and demonstrates significant advantages on MAE, $wF_\beta$ and $S_m$ measures. It suggests that our method has fewer false predictions and better overall performance and that our prediction maps have more similar

TABLE IV
PERFORMANCE ON THE ODI-SOD TEST SUBSET WITH DISCONTINUOUS
EDGE EFFECTS.

| Methods | MAE↓ | $F_\beta$↑ | $wF_\beta$↑ | $S_m$↑ | $E_m$↑ | $maxF$↑ |
|---|---|---|---|---|---|---|
| ZoomNet [10] | 0.071 | 0.801 | 0.752 | 0.797 | 0.869 | 0.860 |
| DDS [7] | 0.074 | 0.778 | 0.733 | 0.79 | 0.875 | 0.837 |
| our | 0.065 | 0.807 | 0.776 | 0.82 | 0.875 | 0.865 |

structures with ground-truth maps. Overall, our method is effective for the discontinuous edge effects in panoramas.

From Fig.13 we can notice that the performances of all the listed methods sharply decrease when processing intricate image samples such as those with very large/small target foreground area ratios, very wide/narrow FoV coverages and severe distortions. However, our method still presents consistent advantages for most image samples and the advantages become more and more evident as the target foreground area or FoV coverage gets smaller and smaller, which indicates that our method is better at processing small targets. On the whole, for samples with different attributes, our method performs better than other methods.

*2) Qualitative Evaluation:* Fig.14 shows some representative results of existing SOTA methods and our method on ODI-SOD test set. We can perceive that most methods fail to segment severely distorted regions well, while our method has more robust adaptability to different distortions (e.g., the first two rows in Fig.14). When processing the object with discontinuous edge effects especially one of the separated parts is very small, other methods usually ignore its part regions, in contrast, our method can segment it more completely (e.g., the middle two rows in Fig.14). For changeable scale objects especially small objects, our method also outperforms other methods (e.g., the last two rows in Fig.14). Generally, our method can obtain better, more complete, more continuous and more uniform segmentation maps than other methods, which is effective for 360° ISOD task.

*3) Complexity Analysis:* In addition to the above evaluations, complexity analysis of the finetuned models is also conducted. We calculate the parameters and MACs (Multiply–Accumulate Operations) of the finetuned models and present the results in Tab.III. From Tab.III, we see that method TRACER [81] has the best parameters and MACs but relatively worse performance, and method ZoomNet [10] has fewer parameters but more MACs. In general, our method obtains a balance between model complexity and performance. For $512 \times 256$ input images, our method runs at a speed of
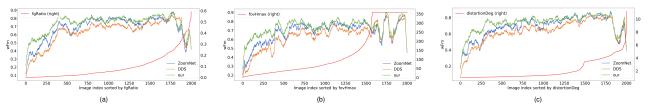
Fig. 13. Performance variation trends with sample attributes including (a) target foreground ratio, (b) max horizontal FoV coverage and (c) distortion degree. In each subfigure, the red line constitutes attribute values and the green line is our trends.
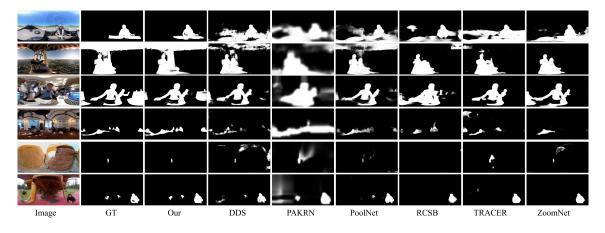


Fig. 14. Visual testing examples of representative state-of-the-art algorithms after finetuning on the ODI-SOD train set.

TABLE V
ABLATION STUDY OF OUR METHOD ON THE ODI-SOD TEST SET. NOTE: HRB, VRB AND ZB ARE THE HORIZONTAL ROTATION, VERTICAL ROTATION AND ZOOMING BRANCHES IN VT.

| Method | HRB | VRB | ZB | SAF | MAE↓ | $F_\beta\uparrow$ | $wF_\beta\uparrow$ | $S_m\uparrow$ | $E_m\uparrow$ | $maxF\uparrow$ |
|--------|-----|-----|-----|-----|------|------|------|------|------|------|
| Baseline | | | | | 0.04 | 0.73 | 0.713 | 0.817 | 0.875 | 0.803 |
| Baseline+VT_H | ✓ | | | | 0.037 | 0.746 | 0.728 | 0.826 | 0.88 | 0.807 |
| Baseline+VT_V | | ✓ | | | 0.038 | 0.739 | 0.719 | 0.822 | 0.876 | 0.805 |
| Baseline+VT_Z | | | ✓ | | 0.038 | 0.742 | 0.725 | 0.824 | 0.878 | 0.811 |
| Baseline+VT_HV | ✓ | ✓ | | | 0.038 | 0.751 | 0.728 | 0.822 | 0.879 | 0.818 |
| Baseline+VT_HZ | ✓ | | ✓ | | 0.037 | 0.751 | 0.731 | 0.827 | 0.88 | 0.816 |
| Baseline+VT_VZ | | ✓ | ✓ | | 0.037 | 0.741 | 0.727 | 0.824 | 0.877 | 0.813 |
| Baseline+VT | ✓ | ✓ | ✓ | | 0.038 | 0.75 | 0.734 | 0.83 | 0.882 | 0.818 |
| Baseline+SAVT(VT+SAF) | ✓ | ✓ | ✓ | ✓ | 0.035 | 0.759 | 0.738 | 0.831 | 0.886 | 0.822 |

7.6 FPS on one GTX 2080Ti GPU.

### D. Ablation Study

To demonstrate the effectiveness of the proposed module SAVT with two sub-modules VT and SAF, we further conduct ablation studies on the ODI-SOD test set. The three branches in VT are also considered for detailed analysis. First, based on the baseline model, we add the gimped VT versions with only a single branch or two branches to test their effectiveness. Then, we try the complete submodel VT and SAF. The experimental results are shown in Tab.V.

*1) Effectiveness of VT:* From the first four rows in Tab.V we observe that all the performances can be improved when only using one transform branch, especially the versions VT_H and VT_Z, which indicates the single transform branch is effective. When randomly adopting two transform branches in VT, most metrics get better than those using only one, suggesting that the models with two transform branches still work well. When utilizing the complete VT, the overall performance

is further enhanced. Compared with the baseline, the MAE score becomes 0.038 from 0.040 and the $F_\beta$ increases to 0.759 from 0.730. It verifies the sub-model VT is effective. It is worth mentioning that more transform branches mean more diversities of features. This requires a stronger feature fusion operation to obtain the desired features. Next, we will further verify the ability of SAVT to fuse different types of transformed features.

*2) Effectiveness of SAF:* SAF is based on VT to assist subsequent feature fusion by adaptively adjusting the weights of different types of transformed features. From Tab.V we find that all the metrics are improved after using SAF. After adding SAF, The MAE score becomes 0.035 from 0.038 and the $F_\beta$ score becomes 0.759 from 0.750, which presents the effectiveness of SAF.

Overall, the performance is significantly improved from the baseline to our final model with the proposed SAVT. As shown in Tab.V, the MAE score decreases to 0.035 from 0.040 and the $F_\beta$ score increases to 0.759 from 0.730. It shows the

TABLE VI
INFLUENCE OF ROTATION DEGREES IN THE HORIZONTAL/VERTICAL
ROTATION BRANCH (HRB/VRB) AND ZOOMING SCALE FACTORS IN THE
ZOOMING BRANCH (ZB). FOR HRB, DEGREES RANGES FROM $-180°$ TO
$180°$ EXCEPT $0°$.

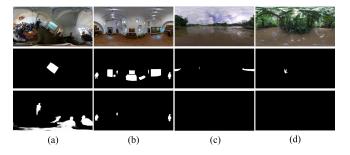| Parameters | MAE↓ | $F_\beta\uparrow$ | $wF_\beta\uparrow$ | $S_m\uparrow$ | $E_m\uparrow$ | $maxF\uparrow$ |
|---|---|---|---|---|---|---|
| Degree step | HRB | | | | | |
| 30° | 0.035 | 0.758 | 0.738 | 0.829 | 0.885 | 0.823 |
| 45° | 0.037 | 0.754 | 0.743 | 0.833 | 0.884 | 0.816 |
| 60° | 0.039 | 0.748 | 0.738 | 0.828 | 0.879 | 0.816 |
| Degrees | VRB | | | | | |
| [±30°] | 0.035 | 0.758 | 0.738 | 0.829 | 0.885 | 0.823 |
| [±60°,±30°] | 0.039 | 0.745 | 0.729 | 0.827 | 0.876 | 0.814 |
| [±45°,±30°] | 0.039 | 0.745 | 0.727 | 0.826 | 0.878 | 0.812 |
| Scale factors | ZB | | | | | |
| [0.8,1.2,0.7,1.3] | 0.035 | 0.758 | 0.738 | 0.829 | 0.885 | 0.823 |
| [0.8,1.2,0.6,1.4] | 0.038 | 0.749 | 0.731 | 0.826 | 0.88 | 0.814 |
| [0.7,1.3,0.5,1.5] | 0.039 | 0.747 | 0.726 | 0.823 | 0.88 | 0.817 |



Fig. 15. Failure cases of our method. Rows from top to bottom: images, ground-truth masks and our results.

proposed module SAVT is very effective for the 360° ISOD task.

*3) Influence of Parameters in VT:* To investigate the influence of transform branches' parameters in VT, We try several groups of parameters about the rotation degrees of the horizontal/vertical branch and the scale factors of the zooming branch based on the final model. From the results in Tab.VI we find that for HRB the overall performance decreases a little when the rotation degree becomes 45° from 30°, and when the degree becomes 60° the performance is more negatively affected. No smaller degree step is tried as the angular resolution is enough for our task (e.g., 1.42 pixels per degree for a 512×256 ERP image). For VRB, we choose ±30° since the extra larger degrees cannot bring better performance. It is reasonable and realistic as the vertical field of view range is ±90° and we usually do not look up or down too much. Moreover, the ERP feature appearance is more sensitive to vertical rotation than horizontal rotation. The inappropriate larger degrees may bring overly deformed appearance and unexpected distractors, which is not beneficial for feature perception and performance gains. As for ZB, the first group parameters are best. Both smaller and bigger scale factors are not suitable. It is also realistic as excessive zooming in/out cannot help to learn better features and appropriate transformations are more important.

### E. Failure Cases

Beyond the successful cases, we show representative failure cases in Fig.15. We find when the scene contains many objects or the target object is camouflaged, our approach is more prone to failure. For example, the scenes in Fig.15(a) and Fig.15(b) contains many objects while our method cannot detect the targets or miss some targets. To some extent, it is caused by the way of defining salient objects as it is hard to define salient targets in panoramic scenes with many objects, especially when the objects have similar saliency. Besides, when target objects are similar with background or camouflaged in background (e.g., the boat in Fig.15(c) or the animal in Fig.15(d)), our method also fails to find the targets.

### F. Discussion

Although our method is effective and outperforms the SOTA methods, it has three limitations. Firstly, the adopted resolution in the study is not large for panoramas with wide FoV and rich information. Small resolutions can lose some important details. Secondly, to use the mature 2D CNNs, the original spherical images must be projected onto the 2D plane, resulting in different degrees of geometrical distortion. Thirdly, the model is not lightweight and efficient enough. Therefore, how to make full use of the original image information and investigate a highly efficient solution need to be further explored in the future. In addition to methods, some special scenes can also be explored. For example, the target objects in the scene sometimes are concealed or camouflaged, "seamlessly" embedded in their surroundings [18], [19], or the objects are in the clutter [82], [83], which are challenging situations for panoramic scenarios and should be further discussed in future work.

## VI. CONCLUSION

In this paper, we construct a 360° omnidirectional image-based SOD dataset, namely ODI-SOD, to explore the salient object detection in panoramic scenes. It has object-level pixel-wise annotations on ERP images and is the largest dataset for 360° ISOD by far to our best knowledge. Moreover, inspired by humans' observing process, we propose a view-aware salient object detection method for 360° ODIs, containing a novel module SAVT with two submodels VT and SAF. VT is designed to simulate the process of looking left and right, up and down, far and near by changing the viewpoint or view distance. SAF aims to adaptively fuse the output features of transform branches and the original learning branch based on different input samples. It can flexibly adjust the weights of different transformed features and obtain better fusion features. Integrated SAVT effectively mitigates the effects of diverse distortion degrees, discontinuous edge effects and changeable object scales. Furthermore, we conduct qualitative and quantitative experiments to explore the proposed method and verify its effectiveness.

## REFERENCES

[1] Y.-C. Su and K. Grauman, "Learning spherical convolution for fast features from 360 imagery," *Advances in Neural Information Processing Systems*, vol. 30, pp. 529–539, 2017.

[2] Y. Zhang, L. Zhang, W. Hamidouche, and O. Deforges, "A fixation-based 360 benchmark dataset for salient object detection," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3458–3462.

[3] K.-T. Ng, S.-C. Chan, and H.-Y. Shum, "Data compression and transmission aspects of panoramic videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 82–95, 2005.

[4] M. Xu, C. Li, S. Zhang, and P. Le Callet, "State-of-the-art in 360 video/image processing: Perception, assessment and compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020.

[5] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in vr applications," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[6] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for hevc-msp," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 155–170, 2017.

[7] J. Li, J. Su, C. Xia, and Y. Tian, "Distortion-adaptive salient object detection in 360° omnidirectional images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 38–48, 2019.

[8] G. Ma, S. Li, C. Chen, A. Hao, and H. Qin, "Stage-wise salient object detection in 360° omnidirectional image via object-level semantical saliency ranking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3535–3545, 2020.

[9] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proceedings. of the AAAI Conference On Artificial Intelligence*, 2021, pp. 1–9.

[10] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," 2022.

[11] "Field of view," https://en.wikipedia.org/wiki/Field_of_view.

[12] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero, "Corners for layout: End-to-end layout recovery from 360 images," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1255–1262, 2020.

[13] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 707–722.

[14] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1155–1162.

[15] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2919–2929.

[16] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[17] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, vol. 7, no. 1, pp. 37–69, 2021.

[18] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[19] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2777–2787.

[20] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[21] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 280–287.

[22] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[23] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[24] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4142–4150.

[25] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.

[26] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2386–2395.

[27] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 186–202.

[28] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[29] G. Li, P. Yan, Y. Xie, G. Wang, L. Lin, and Y. Yu, "Instance-level salient object segmentation," *Computer Vision and Image Understanding*, vol. 207, p. 103207, 2021.

[30] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 205–210.

[31] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1420–1429.

[32] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360 immersive videos," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342.

[33] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360 videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 488–503.

[34] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2190–2195.

[35] Y. Fang, X. Zhang, and N. Imamoglu, "A novel superpixel-based saliency detection model for 360-degree images," *Signal Processing: Image Communication*, vol. 69, pp. 1–7, 2018.

[36] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong, "Spherical criteria for fast and accurate 360 object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 959–12 966.

[37] Y. Zhang, L. Zhang, J. Zhang, K. Wang, W. Hamidouche, and O. Deforges, "Shd360: A benchmark dataset for salient human detection in 360 videos," *arXiv preprint arXiv:2105.11578*, 2021.

[38] Y. Zhang, F.-Y. Chao, G.-P. Ji, D.-P. Fan, L. Zhang, and L. Shao, "Asod60k: Audio-induced salient object detection in panoramic videos," *arXiv preprint arXiv:2107.11629*, vol. 2, no. 3, p. 4, 2021.

[39] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1767–1777, 2021.

[40] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Déforges, "A multi-fov viewport-based visual saliency model using adaptive weighting losses for 360° images," *IEEE Transactions on Multimedia*, vol. 23, pp. 1811–1826, 2020.

[41] Y. Fang, Y. Yao, X. Sui, and K. Ma, "A database for perceived quality assessment of user-generated vr videos," *arXiv preprint arXiv:2206.08751*, 2022.

[42] Y. Fang, L. Huang, J. Yan, X. Liu, and Y. Liu, "Perceptual quality assessment of omnidirectional images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 580–588.

[43] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[44] W. Yongliang, "Cvpr 2019— poolnet: Interpretation of saliency detection papers based on pooling technology cvpr 2019— poolnet: Interpretation of saliency detection papers based on pooling technology."

[45] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[46] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 599–10 606.

[47] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150.

[48] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 025–13 034.

[49] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2311–2318.

[50] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4722–4732.

[51] Z. Zhao, C. Xia, C. Xie, and J. Li, "Complementary trilateral decoder for fast and accurate salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4967–4975.

[52] M. Huang, Z. Liu, G. Li, X. Zhou, and O. Le Meur, "Fanet: Features adaptation network for $360°$ omnidirectional salient object detection," *IEEE Signal Processing Letters*, vol. 27, pp. 1819–1823, 2020.

[53] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 682–690.

[54] K. Yang, X. Hu, and R. Stiefelhagen, "Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 1866–1881, 2021.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[56] F. De Simone, P. Frossard, P. Wilkins, N. Birkbeck, and A. Kokaram, "Geometry-driven quantization for omnidirectional image coding," in *2016 picture coding symposium (PCS)*. IEEE, 2016, pp. 1–5.

[57] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2015, pp. 31–36.

[58] J. Olsen, "The geometry of möbius transformations," *Rochester: University of Rochester*, 2010.

[59] P. Blanchard, "Complex analytic dynamics on the riemann sphere," *Bulletin of the American mathematical Society*, vol. 11, no. 1, pp. 85–141, 1984.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[61] David R. Wilkins, "Möbius transformations and stereographic projection," https://www.maths.tcd.ie/~dwilkins/Courses/MA232A/MA232A_Mich2017/StWrapper.pdf, 2017.

[62] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[64] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 248–255.

[65] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.

[66] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 698–704.

[67] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function," *SCIENTIA SINICA Informationis*, vol. 6, p. 6, 2021.

[68] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604.

[69] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[70] J. Wei, S. Wang, and Q. Huang, "F$^3$net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[71] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton," *IEEE Transactions on Image Processing*, vol. 29, pp. 8652–8667, 2020.

[72] M. Ma12, C. Xia, and J. Li123, "Pyramidal feature shrinking for salient object detection," 2021.

[73] Z. Wu, L. Su, and Q. Huang, "Decomposition and completion network for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6226–6239, 2021.

[74] M.-M. Cheng, S. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[75] S. Yang, W. Lin, G. Lin, Q. Jiang, and Z. Liu, "Progressive self-guided loss for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8426–8438, 2021.

[76] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Scene context-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4156–4166.

[77] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3804–3814, 2021.

[78] J. Li, J. Su, C. Xia, M. Ma, and Y. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Transactions on Image Processing*, vol. 30, pp. 6855–6868, 2021.

[79] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE CVPR*, 2019.

[80] Y. Y. Ke and T. Tsubono, "Recursive contour-saliency blending network for accurate salient object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 2940–2950.

[81] M. S. Lee, W. Shin, and S. W. Han, "Tracer: Extreme attention guided salient object tracing network," *arXiv preprint arXiv:2112.07380*, 2021.

[82] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 186–202.

[83] D.-P. Fan, J. Zhang, G. Xu, M.-M. Cheng, and L. Shao, "Salient objects in clutter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.