Lightweight Distortion-Aware Network for Salient Object Detection in Omnidirectional Images

Mengke Huang, Gongyang Li, Zhi Liu, *Senior Member, IEEE*, and Linchao Zhu, *Member, IEEE*

***Abstract*— Compared with 2D image salient object detection (SOD), SOD in omnidirectional images (or 360° images) usually suffers from geometric distortion. Although existing omnidirectional image SOD (ODI-SOD) methods have improved the detection accuracy obviously, their application may be cumbersome in real scenes due to their high computational cost. To avoid distortion and reduce the computational cost simultaneously in ODI-SOD, we propose a novel lightweight distortion-aware network, named LDNet, in this letter. First, to extract features with less distortion from ODIs, we integrate the distortion-aware convolution and depth-wise separable convolution (DSConv) into distortion-aware DSConv (DDSConv) and replace the regular convolutions in the last two blocks of the ResNet-18 with DDSConvs to obtain our lightweight backbone network (LD-ResNet-18). To enhance spatial information in each channel of the extracted features at each level comprehensively, then, we propose a lightweight distortion-aware channel-wise enhancement (DCE) module (only 0.05M parameters) including DDSConvs with various dilation rates, channel shuffle operation and attention mechanism, and employ a high-to-low dense connection structure to modulate the enhanced multi-level features. Besides, we design a distortion-aware self-correlation (DSC) module (only 0.02M parameters) for mining the contextual dependency of the features via a coarse-fine strategy, and the correlated features are refined by DCE modules and integrated by another dense connection structure. The final saliency map is predicted from the densely integrated features. Compared with 12 state-of-the-art methods on two public datasets, our lightweight LDNet achieves competitive or even better performance with only 2.9M parameters and 3.4G FLOPs, which balances the efficiency and performance.**

***Index Terms*— Lightweight salient object detection, omnidirectional image, distortion-aware and depth-wise separable convolution, distortion-aware self-correlation.**

Manuscript received 13 December 2022; revised 13 February 2023; accepted 25 February 2023. Date of publication 6 March 2023; date of current version 4 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171269 and in part by the China Postdoctoral Science Foundation under Grant 2022M722037. This article was recommended by Associate Editor S. Wang. *(Corresponding author: Gongyang Li.)*

Mengke Huang is with the Shanghai Institute for Advanced Communication and Data Science, and the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with Shanghai Electric Group Co., Ltd., Central Academe, Shanghai 200070, China (e-mail: huangmengke@shu.edu.cn).

Gongyang Li and Zhi Liu are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, and the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: ligongyang@shu.edu.cn; liuzhisjtu@163.com).

Linchao Zhu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: zhulinchao@zju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3253685.

Digital Object Identifier 10.1109/TCSVT.2023.3253685

I. Introduction

SALIENT object detection (SOD), of which the goal is to highlight the visually attended objects or regions in the image, is an essential step for many computer vision tasks such as image segmentation [1], [2], [3], [4], visual tracking [5], image retargeting [6], [7], [8] and image captioning [9]. Many SOD methods [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] in 2D images (TDI-SOD) with the limited field of view, have reached good performances benefiting from convolution neural networks (CNNs). Aiming at omnidirectional images (namely 360° images), which display the spatial information on the 3D spherical surfaces, most existing CNN-based omnidirectional image (ODI) SOD methods [20], [21] explore different schemes to mitigate the geometric distortion caused by projecting ODIs to planes. These ODI-SOD methods with strong backbone CNNs achieve good performances, however, their computation and parameter overheads are heavy. Although several lightweight TDI-SOD methods [22], [23], [24], [25] have been proposed, applying them directly to the ODI-SOD task may lead to suboptimal accuracy due to the lack of solutions for the distortion of equirectangular ODIs.

By replacing the normal convolution with depth-wise separable convolution (DSConv), the parameters of existing lightweight backbone networks [26], [27], [28], [29] for different computer vision tasks are reduced significantly without decreasing the feature representation ability. Similar to lightweight TDI-SOD methods, however, these lightweight backbone networks mainly focus on 2D scenes and neglect the properties of ODIs. Thus, they may be not suitable for ODI-SOD straightly. To avoid the distortion caused by the equirectangular projection in polar regions of the ODIs, previous methods deform the sampling locations of the regular convolution kernels by the gnomonic projection and propose distortion-aware convolution for classification [30], object detection [31] and depth estimation [32] in ODIs.

Inspired by the distortion-aware convolution and DSConv, in this letter, we propose a novel lightweight distortion-aware network, namely LDNet, for ODI-SOD. The proposed LDNet is the first lightweight ODI-SOD method to the best of our knowledge. In LDNet, we lighten the vanilla ResNet-18 [33] with the distortion-aware DSConv (DDSConv) to balance parameters and distorted feature representation capabilities of the network. To capture spatial semantic information efficiently and sufficiently, we design a lightweight distortion-aware channel-wise enhancement (DCE) module via

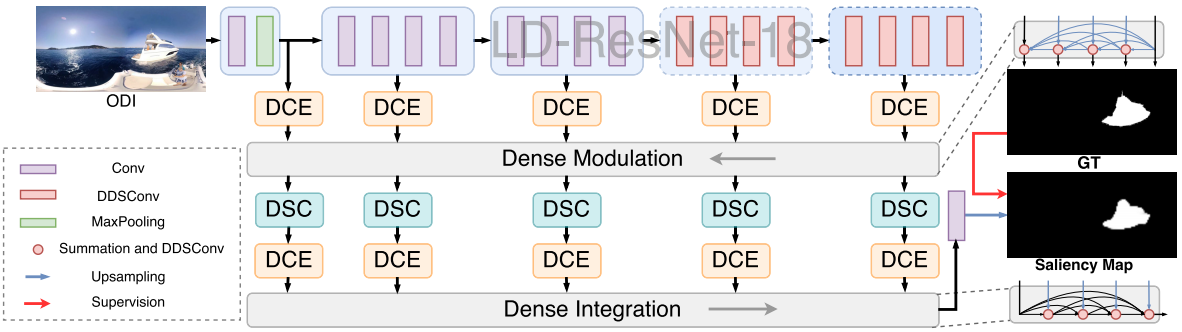

Fig. 1. **The overview of the proposed LDNet**. The output features of lightweight distortion-aware ResNet-18 are first enhanced by the lightweight distortion-aware channel-wise enhancement (DCE) modules at each level. Then, the enhanced multi-level features are modulated by the dense modulation structure, and the densely modulated features are boosted by distortion-aware self-correlation (DSC) modules and DCE modules. Finally, the dense integration structure aggregates multi-level refined features and produces the final saliency map.

DDSConvs with different dilation rates, channel shuffle [28], [29] and attention mechanism [34]. DCE modules are deployed at different levels of the backbone network, and we use a dense connection structure to modulate the feature flow in different semantic levels. Furthermore, we introduce a distortion-aware self-correlation (DSC) module to boost the densely modulated features at multiple levels by a coarse-fine pattern. Finally, the self-correlated features refined by DCE modules are integrated by another reversely dense connection structure, and the final saliency map is predicted from the densely interacted features. In this way, the proposed LDNet achieves competitive performance compared with 12 state-of-the-art CNN-based SOD methods on two ODI-SOD datasets.

The main contributions of this letter are threefold:

- We propose a novel LDNet (only 2.9M parameters) to explore the lightweight ODI-SOD for the first time. In LDNet, we introduce DDSConv to lighten the network and mitigate the distortion synchronously.
- We propose a lightweight distortion-aware channel-wise enhancement module to mine the distorted multi-scale semantic information in each channel of features of ODIs.
- To correlate the contextual semantic features of ODIs, we propose a distortion-aware self-correlation module based on a coarse-fine strategy.

The rest of this letter is organized as follows. we review the ODI-SOD methods and the lightweight CNN-based TDI-SOD methods in Sec. II. The proposed LDNet is detailed in Sec. III. Then we present comprehensive experiments results in Sec. IV. Finally, this letter is concluded in Sec. V.

II. Related Work

A. CNN-Based Salient Object Detection in ODIs

Compared with the great progress of TDI-SOD achieved in the past decades [35], CNN-based ODI-SOD has just begun to gain increasing attention with the gradual prevalence of consumer 360° cameras in recent years. Different from TDIs, ODIs on the spheres suffer from noticeable geometric distortion in the polar regions when projecting them to the 2D plane to obtain the equirectangular ODIs, which are suitable for processing by CNNs. Thus, the existing CNN-based ODI-SOD methods have introduced different strategies to mitigate the distortion in the equirectangular projection of ODIs. In [20], the first end-to-end ODI-SOD method is proposed. The core strategy in the method to alleviate the distortion in equirectangular projection is the distortion-adaptive module, which cuts the equirectangular ODIs into multiple blocks and assigns exclusive convolution kernels to different image blocks, respectively. For leveraging the strengths of equirectangular and cubemap ODIs to alleviate the defect of the visual attribute of their each other, in [21], a projection features adaptation module is proposed to select and aggregate the features from equirectangular and cubemap projections adaptively. In [36], a sample adaptive view transformer module, designed for capturing various features under different views of ODIs by different kinds of transformations, is proposed to improve the ability of the feature toleration of distortion, edge effects, and object scales in ODIs. Unlike the above end-to-end ODI-SOD methods, in [37], the stage-wise ODI-SOD method is proposed and divided into multiple sequential tasks. To mitigate the distortion, this method adopts object-level semantical saliency ranking, fine-level salient object localization and pixel-wise saliency refinement to detect salient objects in ODIs.

Through diverse strategies to mitigate distortion, the above CNN-based ODI-SOD methods achieve satisfactory performance. However, their large number of parameters and heavy computational cost are unfriendly to deploy them on practical applications. To this end, we propose LDNet based on lightweight distortion-aware ResNet-18 [33] and two lightweight distortion-aware modules, which improve the capacity of the CNNs to extract the undistorted semantic information from the distorted features while reducing the computational cost significantly.

B. Lightweight CNN-Based Salient Object Detection in TDIs

Lightweight SOD in TDIs is a burgeoning computer vision task which aims to lessen the parameter and computational cost of the CNN-based models and achieve comparable performance to the CNN-based methods [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] with a large number of parameters and computational cost. In [23], a stereoscopically attentive multi-scale module is proposed, and this lightweight method is designed by multi-level and multi-scale learning. Similarly, in [22], a hierarchical visual perception (HVP) module based

on dense connections is proposed, and the method composed of the lightweight HVP modules and residual attention is designed for learning multi-scale contexts effectively. In [24], the method based on lightweight VGG-16 [38] and feature correlation module is proposed for 2D optical remote sensing images SOD, and the dense lightweight refinement block is designed for the coarse saliency map generated by the feature correlation module. In [39], an extremely-downsampled block is proposed to learn a better global view of the whole image and accurately localize the salient objects, and a scale-correlated pyramid convolution is designed for better multi-level feature fusion.

The above lightweight SOD methods focused on the 2D scenes employ depth-wise separable convolution (DSConv) [26], [27], [28], [29] to reduce the number of parameters and computational cost of the CNNs. Due to far less distortion in TDIs, the convolution sampling locations of the normal DSConv can extract the features without geometry distortion from TDIs. Different from TDIs, however, the equirectangular ODIs, which are projected from the sphere surface to the 2D plane, have significant distortion and are unsuitable for feature extraction by the normal DSConvs. To enable the convolution kernels to adjust their spatial sampling patterns according to their different locations on the sphere, the distortion-aware convolution [30], [31], [32], which adapts the spatial sampling locations of standard convolution kernel in the equirectangular ODIs by the gnomonic projection of ODIs on the sphere, extracts the undistorted features from equirectangular ODIs with noticeable geometry distortion. To extract undistorted features from equirectangular ODIs and reduce the amount of parameters and computational cost of the proposed LDNet, inspired by the above insights, we combine DSConv with distortion-aware convolution and propose the Distortion-aware DSConv (DDSConv). The difference between DDSConv and DSConv is that the spatial sampling pattern of depth-wise convolution in DSConv is the same as the spatial sampling pattern of distortion-aware convolution. We employ the DDSConv to lighten the vanilla ResNet-18 as our backbone network. Moreover, two lightweight and effective modules based on the DDSConv, *i.e.*, DCE and DSC, are proposed to enhance multi-scale channel features and mine the correlation of semantic context, respectively.

III. Proposed Method

A. Network Overview

As shown in Fig. 1, the proposed LDNet consists of four main components: a lightweight distortion-aware backbone network (LD-ResNet-18) for feature extraction, DCE modules, DSC modules, and dense modulation and integration for multi-level features. In LD-ResNet-18, we replace the regular convolutions in the last two blocks of vanilla ResNet-18 with the DDSConv to slim the network. To mitigate the distortion in CNN-extracted features of ODI while reducing parameters, the DDSConv is constructed by distorting kernel sampling locations of the 3×3 depth-wise convolution in DSConv. In this way, the amount of parameters of LD-ResNet-18 (2.2M) is only 18.8% of vanilla ResNet-18 (11.7M).

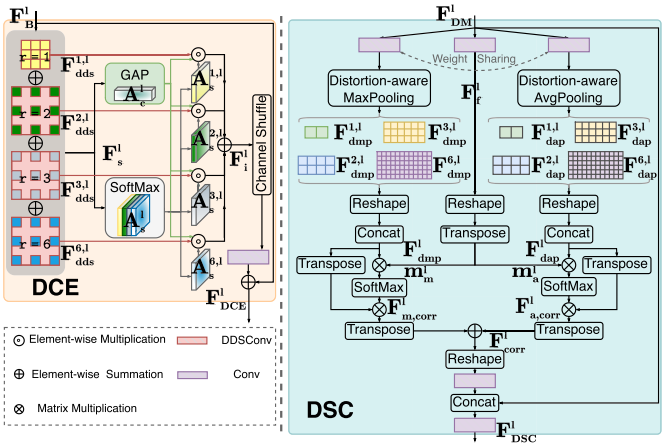

Fig. 2. Structures of the DCE and DSC modules.

Features extracted from LD-ResNet-18 at the l-th level, termed $\{\mathbf{F}_{\mathrm{B}}^{l} \in \mathbb{R}^{c_l \times h_l \times w_l} \mid l \in \{1, 2, 3, 4, 5\}\}$, are first refined by DCE modules for abundant spatial semantic information, where c_l, h_l and w_l are the number of channels, height and width, respectively. To interact with the feature $\mathbf{F}_{\mathrm{B}}^{l}$ enhanced by DCE modules at different levels, we employ a dense multi-level structure inspired by DenseNet [40], [41]. This dense connection structure modulates the low-level features extracted from regular convolutions by high-level features to suit the distortion of ODIs, and it comprises DDSConv, upsampling and element-wise summation at each level, as shown in Fig. 1. Then, we design a DSC module for the multi-level modulated features, and DCE modules are equipped following these DSC modules for feature refinement. Finally, we deploy a dense connection structure with the reverse direction of the previous for integrating multi-level features and predicting the final saliency map.

B. Distortion-Aware Channel-Wise Enhancement Module

Considering the amount of parameters and distortion of features simultaneously, as shown in Fig. 2, we propose a lightweight distortion-aware channel-wise enhancement (DCE) module, which explores spatial features of each channel jointly and regulates channel features adaptively, to refine $\{\mathbf{F}_{\mathrm{B}}^{l} \mid l \in \{1, 2, 3, 4, 5\}\}$ at multiple levels of LD-ResNet-18.

Specifically, we first handle $\mathbf{F}_{\mathrm{B}}^{l}$ with four channel-wise 3×3 DDSConvs with different dilation rates $r \in \{1, 2, 3, 6\}$.For each channel-wise DDSConv, the weights of each channel share the same values. We then integrate $\mathbf{F}_{dds}^{r,l}$ outputted from these dilated DDSConvs by element-wise summation to obtain rich spatial information of each channel. To adaptively enhance the contributory spatial features in each channel, we employ a 3×3 DDSConv to the summated feature $\mathbf{F}_{s}^{l}$, and a SoftMax function along channel dimension is deployed to produce the spatial attention map $\mathbf{A}_{s}^{l} \in [0, 1]^{4 \times h_l \times w_l}$. Each channel of $\mathbf{A}_{s}^{l}$, *i.e.*, $\mathbf{A}_{s}^{r,l} \in [0, 1]^{1 \times h_l \times w_l}$, are broadcasted into $[0, 1]^{c_l \times h_l \times w_l}$, and $\mathbf{A}_{s}^{r,l}$ selectively regulates the spatial features of each channel, which are produced by these four dilated DDSConvs correspondingly. Besides, only focusing on the spatial feature of single channel may

neglect the dependencies of the features between different channels. Thus, we impose a global average pooling layer and a 1×1 convolution layer with ReLU [42] activation function to $\mathbf{F}_s^l$, generating the channel attention map $\mathbf{A}_c^l \in \mathbb{R}^{c_l \times 1 \times 1}$. In this way, the multi-scale features on each channel and the features between different channels in $\mathbf{F}_{\mathrm{B}}^l$ are integrated, *i.e.*,

$$\mathbf{F}_i^l = \sum_{r \in \{1,2,3,6\}} \mathbf{A}_c^l \circledcirc \mathbf{A}_s^{r,l} \circledcirc \mathbf{F}_{dds}^{r,l}, \tag{1}$$

where $\mathbf{F}_i^l$ is the integrated feature and $\circledcirc$ is element-wise multiplication.

To further improve the information flow among different channels in $\mathbf{F}_i^l$ obtained by the above adaptive channel-wise enhancement, we introduce the channel shuffle operation [28], [29] ChanShuf$(\cdot)$ following a 1×1 regular convolution layer $\mathrm{Conv}_{1\times1}(\cdot)$. Eventually, the DCE module produces the final enhanced feature $\mathbf{F}_{\mathrm{DCE}}^l$ by identity mapping as follows:

$$\mathbf{F}_{\mathrm{DCE}}^l = \mathrm{Conv}_{1\times1}(\mathrm{ChanShuf}(\mathbf{F}_i^l)) \oplus \mathbf{F}_{\mathrm{B}}^l, \tag{2}$$

where $\oplus$ is the element-wise summation. Next, the enhanced $\mathbf{F}_{\mathrm{DCE}}^l$ at each level is modulated by the dense modulation structure as shown in Fig. 1, and the multi-level modulated features are denoted as $\{\mathbf{F}_{\mathrm{DM}}^l \in \mathbb{R}^{c_l \times h_l \times w_l} \mid l \in \{1, 2, 3, 4, 5\}\}$.

C. Distortion-Aware Self-Correlation Module

To efficiently exploit the contextual dependencies of the distorted features, we propose a distortion-aware self-correlation module via correlating the fine-grained features with coarse-grained features at each level, as shown in Fig. 2. The fine-grained feature (obtained in the middle part of the DSC module in Fig. 2) keeps the original resolution, and we adopt the adaptive distortion-aware max-/avg-pooling (the left/right part of the DSC module in Fig. 2) to alleviate the distortion in generating the coarse-grained features with reduced resolution. The coarse-fine self-correlation balances the computational cost and feature relevancy, and the parallel max/avg-pooling is deployed for a more robust feature correlation. As shown in Fig. 2, since the processes of the left and right parts in the DSC module are similar, we elaborate on the left part (*i.e.*, the Distortion-aware MaxPooling part) below.

Concretely, we first apply a regular 1×1 convolution layer on the inputted feature $\mathbf{F}_{\mathrm{DM}}^l$ of the DSC module to reduce its channel number from c_l to $\frac{c_l}{4}$. Then, we exploit a multi-scale scheme in distortion-aware max-pooling to obtain sufficient coarse-grained features $\{\mathbf{F}_{dmp}^{p,l} \in \mathbb{R}^{\frac{c_l}{4} \times p \times 2p} \mid p \in \{1, 2, 3, 6\}\}$, which will be reshaped to $\mathbb{R}^{\frac{c_l}{4} \times 2p^2}$ and concatenated together to produce $\mathbf{F}_{dmp}^l \in \mathbb{R}^{\frac{c_l}{4} \times 2(1^2+2^2+3^2+6^2)}$. Here, we implement the distortion-aware max-pooling by deforming the regular adaptive max-pooling sampling locations like the deformation in DDSConv. The fine-grained feature $\mathbf{F}_f^l$, obtained by reducing the channel number of $\mathbf{F}_{\mathrm{DM}}^l$ to $\frac{c_l}{4}$ via another regular 1×1 convolution layer, is reshaped directly and transposed to $\mathbb{R}^{(h_l w_l) \times \frac{c_l}{4}}$. We obtain the coarse-fine self-correlation matrix $\mathbf{m}_m^l$ via matrix multiplication $\otimes$, which calculates the relevance between fine-grained and coarse-grained features, *i.e.*,

$$\mathbf{m}_m^l = (\mathrm{Rsp}(\mathbf{F}_f^l))^{\mathrm{T}} \otimes \mathbf{F}_{dmp}^l, \tag{3}$$

where $\mathrm{Rsp}(\cdot)$ and T are reshaping and transposing operations, respectively.

We exploit SoftMax function to normalize $\mathbf{m}_m^l$ along its rows and columns together, and conduct $\otimes$ between the normalized correlation matrix and multi-scale coarse-grained $\mathbf{F}_{dmp}^l$ after T. The correlated feature $\mathbf{F}_{m,corr}^l$, which depends on multi-scale salient features, can be captured by

$$\mathbf{F}_{m,corr}^l = \mathrm{SoftMax}(\mathbf{m}_m^l) \otimes (\mathbf{F}_{dmp}^l)^{\mathrm{T}}. \tag{4}$$

Similar to the left part, the right part of DSC module (*i.e.*, the Distortion-aware AvgPooling part) produces another correlated feature $\mathbf{F}_{a,corr}^l$, which represents the correlation between the fine-grained features and the coarse-grained comprehensive features via distortion-aware average pooling. The coarse-fine self-correlated feature $\mathbf{F}_{corr}^l$ is obtained by imposing element-wise summation to the transposed $\mathbf{F}_{m,corr}^l$ and $\mathbf{F}_{a,corr}^l$. We reshape $\mathbf{F}_{corr}^l$ to $\mathbb{R}^{\frac{c_l}{4} \times h_l \times w_l}$ and employ a 1×1 regular convolution layer to recover the channel of $\mathbf{F}_{corr}^l$ from $\frac{c_l}{4}$ to c_l. Lastly, the channel-recovered $\mathbf{F}_{corr}^l \in \mathbb{R}^{c_l \times h_l \times w_l}$ and $\mathbf{F}_{\mathrm{DM}}^l$ are concatenated and fused via another 1×1 regular convolution layer, which decreases the number of channels from $2c_l$ to c_l. The feature outputted from the DSC module is

$$\mathbf{F}_{\mathrm{DSC}}^l = \mathrm{Conv}_{1\times1}(\mathrm{Cat}(\mathrm{Conv}_{1\times1}(\mathrm{Rsp}(\mathbf{F}_{corr}^l)), \mathbf{F}_{\mathrm{DM}}^l)), \tag{5}$$

where $\mathrm{Cat}(\cdot)$ is the concatenation operation. We equip the DCE module after the DSC module at each level for feature refinement. Another dense connection structure, which has the reverse fusion direction with the previous dense modulation, is deployed for integrating the features at each level as shown in Fig. 1. The final saliency map $\mathbf{S}$ is predicted from the densely integrated features by a regular 1×1 convolution layer with sigmoid activation function and upsampling.

D. Implementation Details

To train the proposed LDNet efficiently, we combine binary cross-entropy loss, intersection over union loss and SSIM loss [11] as the total loss function of saliency supervision.

We train and test our LDNet on the PyTorch [43] platform with an NVIDIA Titan RTX GPU (24G memory). The weights of the first three blocks of LD-ResNet-18 are initialized by the pre-trained ResNet-18 [33] model on ImageNet [44], and other newly added DDSConvs and 1×1 regular convolution layers are initialized by the normal distribution proposed in [45]. In the training phase, we resize the input size of ODIs to 672×336 and adopt the Adam optimization strategy [46]. We set the training batch size to 8 and the initial learning rate to 0.001. Besides, we adopt the 'poly' policy described in [47] to adjust the learning rate. The end-to-end training converges $\sim$50 epochs, and our code is available at https://github.com/DreaMKHuang/LDNet.git.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: We train and test our LDNet on **360-SSOD** and **360-SOD** datasets, respectively. In 360-SSOD [37], there are 1,105 ODIs with corresponding ground truths, which

TABLE I

QUANTITATIVE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER STATE-OF-THE-ART METHODS.↑ (↓) INDICATES THE LARGER (SMALLER) IS BETTER.THE TOP THREE RESULTS ARE MARKED IN **RED**, **BLUE** AND **GREEN** ACCORDINGLY

Methods	Type	Input size	Speed (fps) ↑	#Param (M) ↓	FLOPs (G) ↓	360-SSOD [37]				360-SOD [20]			
						$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^a\uparrow$	$E_\xi^a\uparrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^a\uparrow$	$E_\xi^a\uparrow$
R3Net [10]	*T.*	300 × 300	2	56.2	47.5	0.741	0.036	0.489	0.742	**0.821**	**0.022**	**0.650**	0.847
BASNet [11]	*T.*	256 × 256	25	87.1	254.8	0.726	0.033	**0.563**	**0.831**	0.790	0.026	**0.657**	0.849
MLFI-MSFF [12]	*T.*	321 × 321	11	48.5	100.9	0.716	0.043	0.398	0.642	0.782	0.033	0.454	0.668
CPDNet [13]	*T.*	352 × 352	62	29.2	59.5	**0.746**	0.034	0.478	0.719	0.795	0.024	0.631	0.851
GCPANet [15]	*T.*	320 × 320	23	67.1	54.3	**0.761**	**0.032**	0.538	0.788	**0.803**	0.024	0.603	0.817
MINet [14]	*T.*	320 × 320	12	47.6	146.3	0.741	**0.029**	**0.582**	**0.837**	0.776	0.025	0.641	0.840
DDS [20]	*O.*	512 × 256	17	27.2	56.3	0.657	0.053	0.436	0.730	0.801	**0.023**	0.637	**0.852**
FANet [21]	*O.*	1024 × 512	4	25.4	2926.2	0.722	**0.032**	0.550	0.806	**0.826**	**0.021**	**0.700**	**0.883**
SAMNet [23]	*L.T.*	336 × 336	44	1.3	0.5	0.725	0.042	0.456	0.697	0.757	0.033	0.509	0.754
HVPNet [22]	*L.T.*	336 × 336	26	1.2	1.1	**0.742**	0.039	0.448	0.700	0.768	0.029	0.521	0.752
CorrNet [24]	*L.T.*	256 × 256	100	4.1	21.1	0.629	0.037	0.484	0.731	0.698	0.031	0.604	0.815
EDN [39]	*L.T.*	384 × 384	28	1.8	4.4	0.694	0.045	0.510	0.815	0.726	0.042	0.542	0.830
Ours	*L.O.*	672 × 336	10	2.9	3.4	0.727	0.035	**0.557**	**0.840**	0.768	0.029	0.617	**0.858**

T.: SOD method for 2D images. *O.*:SOD method for ODIs.
L.T.: lightweight SOD method for 2D images. *L.O.*: lightweight SOD method for ODIs.

contain 850 ODIs for training and 255 ODIs for testing. In 360-SOD [20], there are 500 ODIs with corresponding ground truths, including 400 training ODIs and 100 testing ODIs.

2) Evaluation Metrics: For evaluating the proposed LDNet, we employ six evaluation metrics to evaluate our method and other compared methods, including S-measure ($S_\alpha, \alpha = 0.5$) [48], adaptive F-measure ($F_\beta^a, \beta^2 = 0.3$) [49], adaptive E-measure (E_ξ^a) [50] and Mean Absolute Error ($\mathcal{M}$) for detection accuracy and parameter amount (#Param) and floating point operations per second (FLOPs) with a batch size of 1 (without I/O time) in inference stage.

B. Comparison With State-of-the-Art Methods

1) Compared Methods: We compare our method with 12 state-of-the-art ODI/TDI SOD methods comprehensively, including six large CNN-based methods for 2D SOD (R3Net [10], BASNet [11], MLFI-MSFF [12], CPDNet [13], GCPANet [15], and MINet [14]), two large CNN-based ODI-SOD methods (DDS [20] and FANet [21]) and four lightweight 2D SOD methods (SAMNet [23], HVPNet [22], CorrNet [24] and EDN [39]). For a fair comparison, we utilize CNN-based SOD methods with their default parameters and re-train them on the training sets of 360-SSOD and 360-SOD, respectively, like our method.

2) Computational Complexity and Quantitative Comparison.: We report #Param, FLOPs, S_α, F_β^a, E_ξ^a and $\mathcal{M}$ comparison of our method and the compared methods in Tab. I. Compared with CNN-based SOD methods, the #param and FLOPs of our method are much smaller than them still with competitive performance on 360-SSOD and 360-SOD datasets. Compared with two lightweight methods, *i.e.*, SAMNet and HVPNet, the #param and FLOPs of our method are slightly inferior. However, the quantitative performance of our method on the two datasets significantly outperforms these two methods. As the first lightweight ODI-SOD method, our LDNet is efficient and promising.

3) Visual Comparison: We present the visual comparisons of the investigated methods in Fig. 3. As shown in the 1^{st} and 2^{nd} columns of Fig. 3, most methods can predict the salient objects in relatively easy scenes accurately. In the 3^{rd} and 4^{th} columns of Fig. 3, our method detects more accurate salient objects and eliminates ambiguous objects. In the last two columns of Fig. 3 with cluttered backgrounds, our method detects salient objects correctly and suppresses cluttered backgrounds simultaneously. Overall, our method shows stably superior detection ability in different scenes.

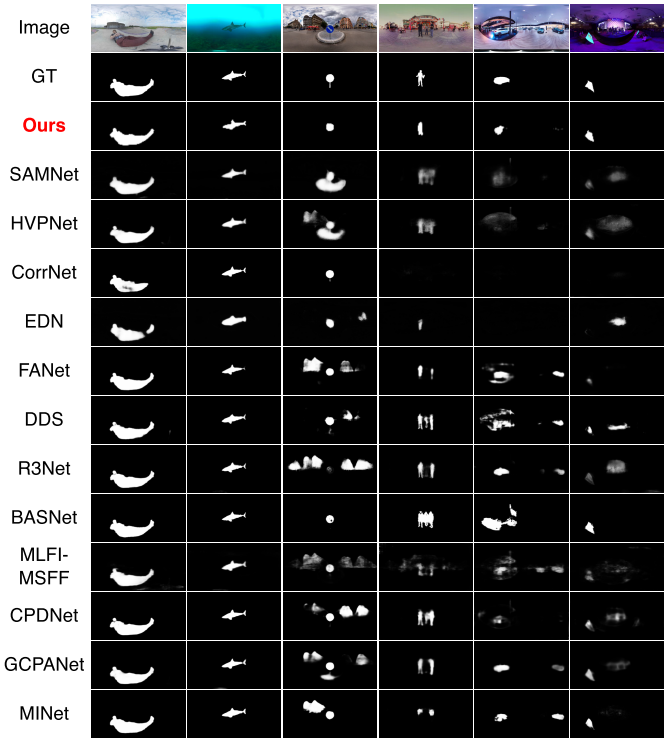

Fig. 3. Visual comparisons of our LDNet with 12 CNN-based SOD methods.

C. Ablation Studies

We provide comprehensive ablation studies to verify the contribution of *1)* the effectiveness of the DCE module, *2)* the importance of the DSC module, *3)* the necessity of the dense modulation structure and dense integration structure, and *4)* the rationality of the proposed DDSConv. Each variant experiment is rigorously re-trained with the same parameter settings and datasets as in Sec. III-D and Sec. IV-A.

We report the computational complexity and quantitative performance of all these variants in Tab. II. Specifically, in variant *w/o DCE*, we remove the DCE modules from each level to illustrate the effectiveness of the DCE module. To evaluate the importance of the DSC module, in *w/o DSC*, the output features of dense modulation structure at each level are refined by DCE modules directly. We present the quantitative performance of *w/o DM*, which is modified by removing the dense modulation structure from the LDNet. In *w/o DI*, we replace the dense integration structure with upsampling and element-wise summation. To prove the rationality of the DDSConv, we replace the DDSConvs in LDNet with DSConv to obtain the variant *w/ DSConv*.

Although minimal computational cost increased, as shown in Tab. II, enhancing distortion-aware channel-wise features at each level, refining features at each level by distortion-aware self-correlation based on coarse-fine strategy, modulating and

TABLE II
ABLATION STUDY FOR THE PROPOSED LDNET

Methods	#Param (M) ↓	FLOPs (G) ↓	360-SSOD [37]				360-SOD [20]			
			$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta^a \uparrow$	$E_\xi^a \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta^a \uparrow$	$E_\xi^a \uparrow$
Ours	2.9	3.4	0.727	0.035	0.557	0.840	0.768	0.029	0.617	0.858
w/o DCE	2.4	3.3	0.716	0.036	0.542	0.825	0.751	0.031	0.605	0.846
w/o DSC	2.8	3.3	0.720	0.036	0.544	0.818	0.755	0.030	0.593	0.848
w/o DM	2.8	3.4	0.722	0.038	0.557	0.831	0.765	0.031	0.608	0.856
w/o DI	2.8	3.3	0.719	0.036	0.536	0.818	0.725	0.032	0.539	0.810
w/ DSConv	2.9	3.4	0.720	0.041	0.545	0.826	0.759	0.031	0.591	0.838

integrating multi-level features densely, and the DDSConvs are effective and rational for the lightweight ODI-SOD.

D. Discussion

Here, we discuss the weaknesses of our method and our future work. We summarize the weaknesses as follows: 1) due to the convolution sampling pattern of the DDSConv needing to be computed according to the latitude and longitude of its locations on the sphere, compared with other methods adopting the uniform convolution sampling locations as shown in Tab. I, the running speed of our method is slowed down, and 2) equirectangular ODIs introduce image borders and damage the continuity of the objects in ODIs on the sphere factitiously, which falsifies the final saliency maps predicted by our trained method. Hence, in future works, we will work on the following two directions: 1) we will speed up the convolution sampling pattern of DDSConv in a more efficient way, and 2) we will introduce equirectangular ODIs projected by different angles into our method to further alleviate the discontinuities in the generated saliency map.

V. CONCLUSION

In this letter, we propose a novel and efficient lightweight framework, LDNet, for ODI-SOD. In LDNet, we first employ distortion-aware DSConvs (DDSConv) and lighten the vanilla ResNet-18 by replacing the regular 3×3 convolutions in the last two blocks with DDSConvs. Then, we propose a lightweight distortion-aware channel-wise feature enhancement module to mine more efficient spatial semantic features at each level. To explore the contextual correlation among features at each level, besides, we introduce an efficient distortion-aware self-correlation module via a coarse-fine strategy. Finally, to integrate multi-level features effectively and predict the final saliency map accurately, we employ dense modulation and integration structures at different stages in the LDNet. Comprehensive experiments demonstrate that our LDNet, only with 2.9M parameters and 3.4G FLOPs, is competitive to state-of-the-art large CNN-based methods and outperforms the lightweight SOD methods significantly on two ODI-SOD datasets.

REFERENCES

[1] G. Li et al., "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Trans. Image Process.*, vol. 30, pp. 1461–1475, 2020.

[2] H. Zhou, L. Yang, X. Xie, and J. Lai, "Selective intra-image similarity for personalized fixation-based object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7910–7923, Nov. 2022.

[3] Z. Tan et al., "Real time video object segmentation in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 175–188, Jan. 2021.

[4] W. Wang, J. Shen, H. Sun, and L. Shao, "Video co-saliency guided co-segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1727–1736, Aug. 2018.

[5] X. Wang et al., "Dynamic attention guided multi-trajectory analysis for single object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4895–4908, Dec. 2021.

[6] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[7] W. Tan, B. Yan, and C. Lin, "Beyond visual retargeting: A feature retargeting approach for visual recognition and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3154–3162, Nov. 2017.

[8] Y. Niu, S. Zhang, Z. Wu, T. Zhao, and W. Chen, "Image retargeting quality assessment based on registration confidence measure and noticeability-based pooling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 972–985, Mar. 2021.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2, pp. 1–21, May 2018.

[10] Z. Deng et al., "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[11] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7471–7481.

[12] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, Oct. 2019.

[13] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3902–3911.

[14] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9410–9419.

[15] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 10599–10606.

[16] L. Zhu et al., "Aggregating attentional dilated features for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3358–3371, Oct. 2020.

[17] L. Sun, Z. Chen, Q. M. J. Wu, H. Zhao, W. He, and X. Yan, "AMPNet: Average- and max-pool networks for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4321–4333, Nov. 2021.

[18] H. Mei et al., "Exploring dense context for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1378–1389, Mar. 2021.

[19] Z. Tu, Y. Ma, C. Li, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 582–593, Feb. 2021.

[20] J. Li, J. Su, C. Xia, and Y. Tian, "Distortion-adaptive salient object detection in 360° omnidirectional images," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 38–48, Jan. 2020.

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

[22] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Lightweight salient object detection via hierarchical visual perception learning," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4439–4449, Sep. 2021.

[23] 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 Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021.

[24] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617712.

[25] G. Li, Z. Liu, X. Zhang, and W. Lin, "Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601111.

[26] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[29] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 122–138.

[30] Q. Zhao, C. Zhu, F. Dai, Y. Ma, G. Jin, and Y. Zhang, "Distortion-aware CNNs for spherical images," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1198–1204.

[31] B. Coors, A. P. Condurache, and A. Geiger, "SphereNet: Learning spherical representations for detection and classification in omnidirectional images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 525–541.

[32] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 732–750.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[35] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.

[36] J. Wu, C. Xia, T. Yu, and J. Li, "View-aware salient object detection for 360° omnidirectional image," *IEEE Trans. Multimedia.*, 2022, doi: 10.1109/TMM.2022.3209015.

[37] 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 Trans. Vis. Comput. Graphics*, vol. 26, no. 12, pp. 3535–3545, Dec. 2020.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[39] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "EDN: Salient object detection via extremely-downsampled network," *IEEE Trans. Image Process.*, vol. 31, pp. 3125–3136, 2022.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[41] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies," *Frontiers Inf. Technol. Electron. Eng.*, vol. 22, no. 12, pp. 1551–1558, 2021.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.

[43] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2019, pp. 8026–8037.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[47] L. C. Chen, G. Papandreou, and I. Kokkinos, "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, Jun. 2017.

[48] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4558–4567.

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

[50] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.