FANet: Features Adaptation Network for 360° Omnidirectional Salient Object Detection

Mengke Huang, Zhi Liu, *Senior Member, IEEE*, Gongyang Li, Xiaofei Zhou, and Olivier Le Meur

***Abstract*—Salient object detection (SOD) in 360° omnidirectional images has become an eye-catching problem because of the popularity of affordable 360° cameras. In this paper, we propose a Features Adaptation Network (FANet) to highlight salient objects in 360° omnidirectional images reliably. To utilize the feature extraction capability of convolutional neural networks and capture global object information, we input the equirectangular 360° images and corresponding cube-map 360° images to the feature extraction network (FENet) simultaneously to obtain multi-level equirectangular and cube-map features. Furthermore, we fuse these two kinds of features at each level of the FENet by a projection features adaptation (PFA) module, for selecting these two kinds of features adaptively. Finally, we combine the preliminary adaptation features at different levels by a multi-level features adaptation (MLFA) module, which weights these different-level features adaptively and produces the final saliency maps. Experiments show our FANet outperforms the state-of-the-art methods on the 360° omnidirectional SOD datasets.**

***Index Terms*—360° omnidirectional image, salient object detection, equirectangular and cube-map projection, projection features adaptation, multi-level features adaptation.**

I. Introduction

SALIENT object detection (SOD), which aims to capture the most visually attractive objects in an image, is an underlying vision problem and plays an important role in a wide range of applications such as image/video segmentation [1]–[6], image retargeting [7] and visual tracking [8], [9]. Conventional image SOD models [10]–[15] have reached good performance in the limited field-of-view (FoV) scenes with the rapid development of the convolutional neural networks (CNNs). Due to the sphere-to-plane projection distortion, however, adopting these conventional CNNs based SOD models directly to 360° omnidirectional images, which exhibit the real 3D world, may fail to highlight salient objects in these images.

Equirectangular projection (ERP) [16] is one of the most common methods for storing 360° omnidirectional images as standard 2D images. The equirectangular 360° images display the global object information on the 2D plane, but in these images, the undesirable distortion caused by the sphere-to-plane projection falsifies the real semantic information. Although several non-CNN algorithms [17]–[19] have been proposed to cope with inappropriate distortions, most existing CNNs based SOD models may not be able to highlight precise salient objects from the distorted semantic information because CNNs are sensitive to the regular-grid data instead of the distorted data [20]. To deal with this distortion in equirectangular 360° images, Li *et al.* [21] proposed a distortion-adaptive module which convolves different image blocks with different convolutional kernels to overcome the distortion mentioned above. Nevertheless, these image blocks obtained by cutting equirectangular 360° images directly may not conform with the real 360° omnidirectional scenes.

Compared with ERP, cube-map projection (CMP) [16] by dividing a 360° omnidirectional image into six faces of a cube introduces less geometric distortion, and this projection is an intuitively approximate representation of real 360° scenes. Thus, the cube-map images, which are mapped from equirectangular 360° images by the equirectangular to cube-map projection (E2C), can be treated as semantically-related but less-distorted augmented data for training, and these images are approximated as regular-grid data which can be processed by CNNs more accurately. In this perspective, although it contains fewer data compared with most standard 2D image SOD benchmark datasets [22]–[25], the 360-SOD [21] dataset, which is the largest dataset for 360° omnidirectional image SOD at present, can be augmented by E2C during training. As mentioned in [26], however, CMP may lead to the discontinuities of objects on the boundaries of the cube's faces.

For leveraging the respective advantages of the features, which are extracted from equirectangular and cube-map 360° images by CNN, to overcome the deficiency caused by only exploiting one of the above two types of features, we propose two features adaptation modules for fusing the features of equirectangular and cube-map 360° images adaptively at each level of CNN and then integrating the fused multi-level features adaptively for detecting salient objects in 360° omnidirectional images robustly. The main contributions of our work are three-fold:

1) We propose a novel end-to-end Features Adaptation Network (FANet), whose inputs are the 2D equirectangular 360° image and its corresponding six cube-map images, for 360° omnidirectional image SOD.
2) We propose a Projection Features Adaptation (PFA) module to select and fuse features extracted from equirectangular and cube-map 360° images adaptively, and we equip the PFA module to each level of backbone feature extraction network.

Manuscript received August 5, 2020; revised September 23, 2020; accepted September 26, 2020. Date of publication October 2, 2020; date of current version October 21, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61771301 and Grant 61901145. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sheng Li. (*Corresponding author: Zhi Liu.*)

Mengke Huang, Zhi Liu, and Gongyang Li are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: huangmengke@shu.edu.cn; liuzhisjtu@163.com; ligongyang@shu.edu.cn).

Xiaofei Zhou is with School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: zxforchid@outlook.com).

Olivier Le Meur is with IRISA, University of Rennes 1, 35042 Rennes, France (e-mail: olemeur@irisa.fr).

Digital Object Identifier 10.1109/LSP.2020.3028192

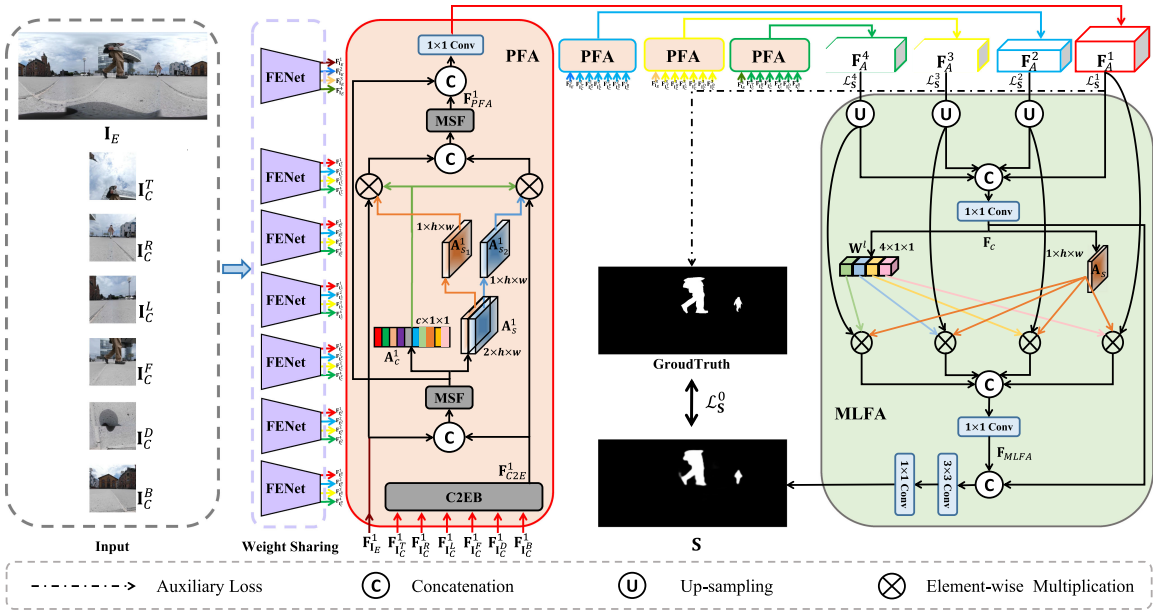

Fig. 1. The overview of our FANet. The weight-shared feature extraction network (FENet) is responsible for extracting multi-level features of input images. Next, the features of equirectangular and cube-map 360° images are fused in the projection features adaptation (PFA) modules at each level. Finally, the multi-level features adaptation (MLFA) module combines the fused features to generate the final saliency map.

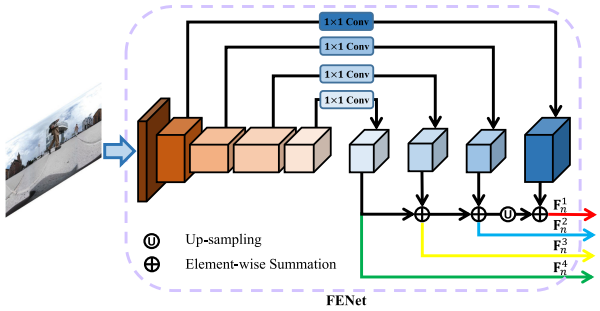

Fig. 2. The structure of the FENet. The input channels of above 1×1 convolutional layers from top to bottom are 256, 512, 1024 and 2048, respectively. Their output channels are 128.

3) We propose a Multi-Level Features Adaptation (MLFA) module to weight and integrate features generated from PFA modules selectively in order to produce the high-quality saliency maps.

II. Proposed Method

A. Feature Extraction Network

We denote the 2D equirectangular 360° image as $\mathbf{I}_E \in \mathbb{R}^{3\times H\times W}$, and the corresponding 90°-FoV cube-map representation $\mathbf{I}_C$ can be transformed by the equirectangular to cube-map (E2C) projection $\mathcal{P}_{E\to C}$. In fact, $\mathbf{I}_C$ is the set of the six planes $\{\mathbf{I}_C^B, \mathbf{I}_C^D, \mathbf{I}_C^F, \mathbf{I}_C^L, \mathbf{I}_C^R, \mathbf{I}_C^T\}$, and each plane $\mathbf{I}_C^i \in \mathbb{R}^{3\times \mathbf{w}\times \mathbf{w}}$ $(i \in \{B, D, F, L, R, T\})$ represents the back, down, front, left, right and top faces of the cube respectively, where $\mathbf{w}$ is the edge length of the cube. For extracting features of $\mathbf{I}_E$ and $\mathbf{I}_C$, we input these images to the feature extraction network (FENet) $\mathcal{N}_{FE}$ presented in Fig. 2 with the shared weights. In $\mathcal{N}_{FE}$, we adopt ResNet [27] based feature pyramid network (FPN) [28], which is a valid CNN structure for object detection. In addition, as the DeepLab [29] described, we set the dilation rates in the backbone ResNet's last two residual blocks to two for larger receptive field. Hence, the outputs of $\mathcal{N}_{FE}$ at each level can be represented by $\mathbf{F}_n^l = \mathcal{N}_{FE}(\mathbf{I}_E, \mathbf{I}_C)(n \in \{\mathbf{I}_E, \mathbf{I}_C\}, l \in \{1, 2, 3, 4\})$, and the channel of $\mathbf{F}_n^l$ is 128.

Compared with $\mathbf{F}_{\mathbf{I}_E}^l$, $\mathbf{F}_{\mathbf{I}_C}^l$ include more plausible spatial detailed information because the less-distorted cube-map 360° images are appropriate for inferring by FENet.

B. Projection Features Adaptation Module

For utilizing the global object information and less-distorted details in the features of equirectangular and cube-map 360° images, we design the projection features adaptation (PFA) module through attention mechanism to select and fuse the features $\mathbf{F}_n^l$ at each level adaptively as shown in Fig. 1.

Specifically, we denote inverse projection of $\mathcal{P}_{E\to C}$, cube-map to equirectangular (C2E) projection, is $\mathcal{P}_{C\to E}$, which can be extended to high-dimensional tensors. Next, the features $\mathbf{F}_{\mathbf{I}_C}^l$ are fed into the C2E projection block (C2EB), which includes $\mathcal{P}_{C\to E}$ and a following 3×3 convolutional layer. The C2EB is for re-projecting the cube-map features containing more reliably spatial details to $\mathbf{F}_{C2E}^l$ which match the spatial sizes of $\mathbf{F}_{\mathbf{I}_E}^l$, *i.e.*$\mathbf{F}_{C2E}^l = Conv_{3\times 3}(\mathcal{P}_{C\to E}(\mathbf{F}_{\mathbf{I}_C}^l))$, where $Conv_{3\times 3}(\cdot)$ represents the 3×3 convolutional layer. Hence, $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ can be concatenated and fed into a multi-scale fusion (MSF) block which is similar with the atrous spatial pyramid pooling (ASPP) module in [29] for capturing multi-scale spatial information, whereas the difference between ASPP and MSF is that we add a 1×1 convolutional layer $Conv_{1\times 1}(\cdot)$ before the four parallel dilated convolutional layers to reduce the concatenated channel from 256 to 128.

To emphasize the global and detailed spatial features in $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ adaptively, we exploit the fused multi-scale features $\mathbf{F}_f^l = MSF(Concat(\mathbf{F}_{\mathbf{I}_E}^l, \mathbf{F}_{C2E}^l))$ to infer the spatial attention

$\mathbf{A}_s^l \in \mathbb{R}^{2\times h\times w}$ by $\mathbf{A}_s^l = \sigma(Conv_{1\times1}(\mathbf{F}_f^l))$, where $Concat(\cdot)$ and $\sigma(\cdot)$ represent *concatenation* operation and channel-wise *Softmax* activation function respectively, and the $Conv_{1\times1}(\cdot)$ reduces the channel from 128 to 2. By this *Softmax* function, the spatial attention scores of each position (x, y) in both channels of $\mathbf{A}_s^l$ add up to 1:

$$\mathbf{A}_{s_1}^l(x,y) + \mathbf{A}_{s_2}^l(x,y) = 1, \tag{1}$$

where $\mathbf{A}_{s_1}^l \in \mathbb{R}^{1\times h\times w}$ and $\mathbf{A}_{s_2}^l \in \mathbb{R}^{1\times h\times w}$ $(0 \le \mathbf{A}_{s_1}^l(x,y), \mathbf{A}_{s_2}^l(x,y) \le 1)$ are the channel features of $\mathbf{A}_s^l$ and adapt $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ by the spatial attention scores.

Simultaneously, through $\mathbf{A}_C^l = \delta(Conv_{1\times1}(Concat(AvgPool(\mathbf{F}_f^l), MaxPool(\mathbf{F}_f^l))))$, the channel attention $\mathbf{A}_c^l \in \mathbb{R}^{c\times1\times1}$ is also inferred, where $\delta(\cdot)$, $AvgPool(\cdot)$ and $MaxPool(\cdot)$ represent *sigmoid* activation function, *average pooling* operation and *maximum pooling* operation respectively, and the $Conv_{1\times1}(\cdot)$ reduces the channel from 256 to 128. Next, we broadcast the channel attention scores along the spatial dimension and vice versa to facilitate computation. After the broadcast, the channel attention scores $\mathbf{A}_c^l$ and the spatial attention scores $\mathbf{A}_{s_1}^l$, $\mathbf{A}_{s_2}^l$, which are learned from the fused multi-scale features $\mathbf{F}_f^l$, multiply $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ respectively and select credible spatial and channel information in $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ adaptively. Furthermore, the multiplied features are fused by the $Concat(\cdot)$ and MSF block, and the feature of PFA is computed as:

$$\begin{aligned}\mathbf{F}_{PFA}^l = MSF(Concat(&\mathbf{A}_c^l \otimes \mathbf{A}_{s_1}^l \otimes \mathbf{F}_{\mathbf{I}_E}^l,\\ &\mathbf{A}_c^l \otimes \mathbf{A}_{s_2}^l \otimes \mathbf{F}_{C2E}^l)),\end{aligned} \tag{2}$$

where $\otimes$ denotes the element-wise multiplication. As Eq. 2, for selecting the global object information and spatial details in $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ adaptively, the spatial features of $\mathbf{F}_{\mathbf{I}_E}^l$ or $\mathbf{F}_{C2E}^l$, which contribute or affect the inference of the model, are retained or suppressed by the element-wise multiplication with higher or lower spatial attention scores ($\mathbf{A}_{s_1}^l$ or $\mathbf{A}_{s_2}^l$). Furthermore, $\mathbf{A}_c^l$, which emphasize the consistent inter-channel relationship of features in $\mathbf{F}_{\mathbf{I}_E}^l$ and $\mathbf{F}_{C2E}^l$ adapted by $\mathbf{A}_{s_1}^l$ and $\mathbf{A}_{s_2}^l$, conduct element-wise multiplication with these adapted features respectively to refine them in channel dimension.

At the end of this module, we concatenate $\mathbf{F}_f^l$ and $\mathbf{F}_{PFA}^l$ and reduce the channel of the concatenated feature from 256 to 128 by a $Conv_{1\times1}(\cdot)$ to obtain the output of PFA module $\mathbf{F}_A^l = Conv_{1\times1}(Concat(\mathbf{F}_f^l, \mathbf{F}_{PFA}^l))$.

C. Multi-Level Features Adaptation Module

As presented in Fig. 1, for combining these four features $\mathbf{F}_A^{1\sim4}$ at different levels selectively, we exploit different weights and a common spatial attention for each feature to adapt and integrate these features in the multi-level features adaptation (MLFA) module, and the final saliency map is generated at the end of MLFA module.

To be more specific, $\mathbf{F}_A^1$ is concatenated with the features, $\mathbf{F}_A^2, \mathbf{F}_A^3$ and $\mathbf{F}_A^4$, which are processed through up-sampling operation $U(\cdot)$ at first, and the channel of concatenated features is reduced from 512 to 128 by a $Conv_{1\times1}(\cdot)$ to get $\mathbf{F}_c$. After that, the separate weights $\mathbf{W}^l \in \mathbb{R}^{4\times1\times1}$ for $\mathbf{F}_A^{1\sim4}$ is learned by $\sigma(\cdot)$, $AvgPool(\cdot)$ and $MaxPool(\cdot)$ on the feature $\mathbf{F}_c$, *i.e.*$\mathbf{W}^l = \sigma(Conv_{1\times1}(Concat(AvgPool(\mathbf{F}_c), MaxPool(\mathbf{F}_c))))$, where $0 \le \mathbf{W}^l \le 1, \sum_{l=1}^4 \mathbf{W}^l = 1$ and the $Conv_{1\times1}(\cdot)$ is for reducing the channel from 256 to 4.

Meanwhile, the common spatial attention map $\mathbf{A}_s \in \mathbb{R}^{1\times h\times w}$ is learned by $\mathbf{A}_s = \delta(Conv_{1\times1}(\mathbf{F}_c))$, where $Conv_{1\times1}(\cdot)$ reduces the channel from 128 to 1. As mentioned in Section II-B, we also broadcast $\mathbf{W}^l$ along the spatial and channel dimension, and spatial attention $\mathbf{A}_s$ is broadcasted along channel dimension. Next, through a $Concat(\cdot)$ and a $Conv_{1\times1}(\cdot)$ for reducing the channel from 512 to 128, the weights $\mathbf{W}^l$ of $\mathbf{F}_A^{1\sim4}$ and the common spatial attention $\mathbf{A}_s$ multiply the corresponding features at different levels to obtain the multi-level adaptive features

$$\begin{aligned}\mathbf{F}_{MLFA} = Conv_{1\times1}(Concat(&\\ &\mathbf{A}_s \otimes \mathbf{W}^1 \otimes \mathbf{F}_A^1, \mathbf{A}_s \otimes \mathbf{W}^2 \otimes U(\mathbf{F}_A^2),\\ &\mathbf{A}_s \otimes \mathbf{W}^3 \otimes U(\mathbf{F}_A^3), \mathbf{A}_s \otimes \mathbf{W}^4 \otimes U(\mathbf{F}_A^4))).\end{aligned} \tag{3}$$

These learnable weights increase or reduce importance of the features at different levels adaptively, *i.e.* the weighted $\mathbf{F}_A^{1\sim4}$ by $\mathbf{W}^l$ are features which are more beneficial for the final prediction. Furthermore, the spatial attention $\mathbf{A}_s$, which keeps the common spatial information of weighted $\mathbf{F}_A^{1\sim4}$ at each level, improves the reliability of these weighted features.

Lastly, $\mathbf{F}_{MLFA}$ and $\mathbf{F}_c$ concatenate with each other, and the final saliency map $\mathbf{S}$ is produced from the concatenated feature after a $Conv_{3\times3}(\cdot)$ for reducing channel from 256 to 128 and a $Conv_{1\times1}(\cdot)$ with a *sigmoid* activation function $\delta(\cdot)$.

D. Implementation Details

We adopt the summation of cross-entropy loss $\mathcal{L}_{CE}$ and mean absolute error loss $\mathcal{L}_{MAE}$ as the loss function $\mathcal{L}_{\mathbf{S}}^j = \mathcal{L}_{CE} + \mathcal{L}_{MAE}(j \in \{0,1,2,3,4\})$. For capturing more precise features, we impose auxiliary loss functions $\mathcal{L}_{\mathbf{S}}^1$, $\mathcal{L}_{\mathbf{S}}^2$, $\mathcal{L}_{\mathbf{S}}^3$ and $\mathcal{L}_{\mathbf{S}}^4$ on $\mathbf{F}_A^{1\sim4}$ for deep supervisions as shown in Fig. 1. Therefore, the total loss function is $\mathcal{L}_{total} = \sum_{j=0}^4 \mathcal{L}_{\mathbf{S}}^j$.

We implement our model by PyTorch [30] framework in a workstation with a NVIDIA Titan RTX GPU (24G memory). The weights of backbone ResNet in FENet are initialized by the pre-trained ResNet-50 [27] model on ImageNet [31], and the normal distribution proposed in [32] is adopted to initialize the weights of newly added convolutional layers. We also resize input equirectangular 360° images to 1024×512, and the size of input cube-map 360° images is 256×256. We utilize stochastic gradient descent (SGD) algorithm for training the model in an end-to-end way. The training batch size is set to 4 and the initial learning rate with momentum 0.9 and weight decay 0.0005 is set to 0.002. More than that, the learning rates of layers except backbone ResNet-50 are set to 10 times larger and the 'poly' learning rate policy described in [33] is adopted. At last, the model converges after 40 epochs (about 3.5h). Our code is available at https://github.com/DreaMKHuang/FANet.git.

III. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate our FANet on two public datasets. $\mathbf{360-SOD}$ [21] consists of 500 equirectangular 360° images including 400 training images and 100 testing images. $\mathbf{F-360iSOD}$ [34] contains 107 equirectangular 360° images.

2) Evaluation Metrics: We utilize the training set of 360-SOD to train our FANet and adopt S-measure (S) [35], mean absolute

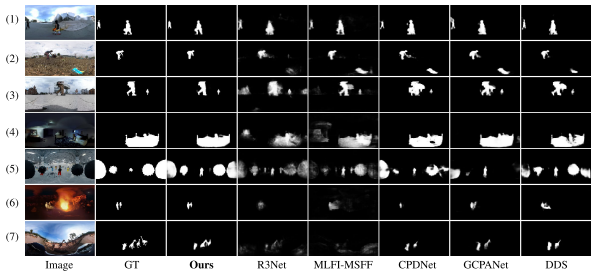

Fig. 3. Visual comparisons to the state-of-the-art methods.

TABLE I
QUANTITATIVE PERFORMANCE OF FANET AND THE STATE-OF-THE-ART RESULTS INCLUDING FIVE METRICS ON THE 360-SOD AND F-360iSOD DATASETS. THE TOP THREE RESULTS ARE SHOWN IN RED, BLUE, AND GREEN. ↑ AND ↓ REPRESENT LARGER AND SMALLER IS BETTER, RESPECTIVELY.

Datasets	Metrics	R3Net	MLFI-MSFF	CPDNet	GCPANet	DDS	**Ours**
360-SOD	$S\uparrow$	0.750	0.782	0.795	0.803	0.799	0.826
	$MAE\downarrow$	0.039	0.033	0.024	0.024	0.023	0.021
	$E\uparrow$	0.661	0.668	0.851	0.817	0.854	0.883
	$F_\beta\uparrow$	0.429	0.454	0.631	0.603	0.638	0.700
	$F_\beta^w\uparrow$	0.454	0.499	0.634	0.628	0.643	0.697
F-360iSOD	$S\uparrow$	0.509	0.580	0.574	0.570	0.577	0.587
	$MAE\downarrow$	0.065	0.081	0.064	0.070	0.061	0.061
	$E\uparrow$	0.705	0.639	0.743	0.718	0.760	0.747
	$F_\beta\uparrow$	0.295	0.300	0.364	0.338	0.364	0.381
	$F_\beta^w\uparrow$	0.116	0.231	0.302	0.276	0.295	0.317

error (MAE) [36], adaptive E-measure (E) [37], adaptive F-measure ($F_\beta, \beta^2 = 0.3$) [38] and weighted F-measure (F_β^w) [39] as the quantitative evaluation metrics.

B. Comparison With the State-of-the-Arts

We compare our model with five state-of-the-art CNN-based SOD models on F-360iSOD and the testing set of 360-SOD. Concretely, these models for comparison include four SOD models for standard 2D images, *i.e.* R3Net [14], MLFI-MSFF [10], CPDNet [12], GCPANet [13], and one SOD model for 360° omnidirectional images, *i.e.* DDS [21]. For a fair comparison, following [21], we fine-tune these four SOD models for standard 2D images on the training set of 360-SOD. As for DDS, we use the saliency maps and the parameters provided by the authors.

1) Qualitative Performance Comparison: We show some typical results in Fig. 3. Specifically, it can be seen that most of the methods perform well in the 1st row of Fig. 3, which has a simple scene. In some cases including ambiguous objects as shown in the 2nd and 3rd rows of Fig. 3, our model can detect accurate objects and suppress inconspicuous objects. In comparison, our model can capture more complete salient objects in 4th and 5th rows of Fig. 3, and our model can highlight small and distorted salient objects more robustly as illustrated in the last two rows of Fig. 3.

2) Quantitative Performance Comparison: We also provide the S-measure, MAE, adaptive E-measure, adaptive F-measure and weighted F-measure of our FANet and the state-of-the-arts on 360-SOD and F-360iSOD in Table I. Although the results of R3Net, MLFI-MSFF, CPDNet and GCPANet are generated by the fine-tuned models on 360-SOD, they cannot segment salient objects precisely due to the lack of solutions for distortion mitigation, and the model DDS designed for equirectangular 360° images is able to perform better than these SOD models for 2D images. Our results demonstrate that incorporating the corresponding less-distorted cube-map images and fusing the equirectangular and cube-map features is effective and better-performing. In addition, the average running time of FANet for processing an equirectangular 360° image with a resolution of 1024×512 is 0.26s.

TABLE II
ABLATION ANALYSIS OF THE PROPOSED FANET ON THE 360-SOD DATASET. THE BEST RESULT IN EACH COLUMN IS **Bold**.

Methods \ Metrics	$S\uparrow$	$MAE\downarrow$	$E\uparrow$	$F_\beta\uparrow$	$F_\beta^w\uparrow$
Ours	**0.826**	**0.021**	**0.883**	**0.700**	**0.697**
w/o CMP	0.820	0.022	0.848	0.662	0.671
w/o ERP	0.722	0.040	0.695	0.450	0.441
w/o PFA	0.825	0.022	0.860	0.667	0.694
w/o MLFA	0.823	0.023	0.855	0.680	0.689
w/o AL	0.799	0.025	0.821	0.633	0.595

C. Ablation Study

In this section, we conduct experiments to assess 1) the importance of fusion features from two different projections, 2) the effectiveness of PFA module and MLFA module, and 3) the influence of auxiliary losses for deep supervision. We change one component each time to assess individual contributions. All the variant models are retrained with the same hyper-parameters and training set as described in Section II-D.

We present the quantitative performance in Table II. Specifically, *w/o CMP* does not include cube-map 360° images in the input data, *w/o ERP* does not include equirectangular 360° images in the input data, *w/o PFA* deletes the projection features adaptation in the PFA module, *w/o MLFA* deletes the MLFA module and *w/o AL* is without the auxiliary losses on the features fused by PFA module.

In more details, because of the lack of features from another different projection, *w/o CMP* and *w/o ERP* do not include the PFA module but still adopt MLFA module to select multi-level features adaptively. To illustrate the effectiveness of projection features adaptation in the PFA module, in *w/o PFA*, we still conduct the MSF block but remove the projection features adaptation in the PFA module. From Table II, we found that selecting and fusing features of equirectangular and cube-map 360° images at each level adaptively, weighting and combining the multi-level fused features and auxiliary losses can further improve the performance.

IV. CONCLUSION

In this letter, we propose a Features Adaptation Network (FANet). The FANet includes two important modules, the projection features adaptation (PFA) module and the multi-level features adaptation (MLFA) module. The PFA module introduces the attention mechanism to emphasize and fuse the features from equirectangular and cube-map projections adaptively at different levels of the feature extraction network (FENet). Next, the features fused by PFA modules are integrated by different adaptive weights for different levels and a common spatial attention, and the final saliency map is generated at the end of the MLFA module. Experimental results demonstrate that our FANet significantly outperforms five state-of-the-arts on the 360° omnidirectional image SOD datasets in terms of five evaluation metrics.

REFERENCES

[1] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 494–10 503.

[2] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8546–8556.

[3] G. Li, Z. Liu, R. Shi, and W. Wei, "Constrained fixation point based segmentation via deep neural network," *Neurocomputing*, vol. 368, pp. 180–187, 2019.

[4] P. Zhang, P. Yan, J. Wu, J. Liu, and F. Shen, "Unsupervised saliency detection in 3-D-video based on multiscale segmentation and refinement," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1384–1388, Sep. 2018.

[5] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, Jan. 2014.

[6] J. Li, S. He, H. Wong, and S. Lo, "Proposal-driven segmentation for videos," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1098–1102, Aug. 2019.

[7] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1511–1518.

[8] Z. Chi, H. Li, H. Lu, and M. Yang, "Dual deep network for visual tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2005–2015, Apr. 2017.

[9] C. Ma, Y. Xu, B. Ni, and X. Yang, "When correlation filters meet convolutional neural networks for visual tracking," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1454–1458, Oct. 2016.

[10] 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, 2019.

[11] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[12] 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.*, 2019, pp. 3902–3911.

[13] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proc. Thirty-Fourth AAAI Conf. Artif. Intell.*, 2020, pp. 10 599–10 606.

[14] Z. Deng *et al.*, "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 684–690.

[15] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, pp. 1–17, 2020.

[16] T. Maugey, O. Le Meur, and Z. Liu, "Saliency-based navigation in omnidirectional image," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process.*, 2017, pp. 1–6.

[17] M. Lourenco, J. P. Barreto, and F. Vasconcelos, "sRD-SIFT: Keypoint detection and matching in images with radial distortion," *IEEE Trans. Robot.*, vol. 28, no. 3, pp. 752–760, Jun. 2012.

[18] A. Furnari, G. M. Farinella, A. R. Bruna, and S. Battiato, "Distortion adaptive sobel filters for the gradient estimation of wide angle images," *J. Vis. Commun. Image Representation*, vol. 46, pp. 165–175, 2017.

[19] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J.-P. Thiran, "Scale invariant feature transform on the sphere: Theory and applications," *Int. J. Comput. Vis.*, vol. 98, no. 2, pp. 217–241, 2012.

[20] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[21] 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.

[22] M.-M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.

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

[24] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3796–3805.

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

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

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

[28] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[29] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[30] A. Paszke *et al.* "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inform. Process. Syst. 32*, 2019, pp. 8026–8037.

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

[32] 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.*, 2015, pp. 1026–1034.

[33] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[34] Y. Zhang, L. Zhang, W. Hamidouche, and O. Deforges, "A fixation-based 360° benchmark dataset for salient object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 3458–3462.

[35] 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.*, 2017, pp. 4558–4567.

[36] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[37] 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.*, 2018, pp. 698–704.

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

[39] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.