Deeper Look at Image Salient Object Detection: Bi-Stream Network With a Small Training Dataset

Zhenyu Wu, Shuai Li, Chenglizhao Chen, Aimin Hao, and Hong Qin

***Abstract*—Compared with the conventional hand-crafted approaches, the deep learning based ISOD (image salient object detection) models have achieved tremendous performance improvements by training exquisitely crafted fancy networks over large-scale training sets. However, do we really need large-scale training set for ISOD? In this article, we provide a deeper insight into the interrelationship between the ISOD performance and the training data. To alleviate the conventional demands for large-scale training data, we provide a feasible way to construct a novel small-scale training set, which only contains 4 K images. To take full advantage of this new set, we propose a novel bi-stream network consisting of two different feature backbones. Benefit from the proposed gate control unit, this bi-stream network is able to achieve complementary fusion status for its subbranches. To our best knowledge, this is the first attempt to use a small-scale training set to compete with other large-scale ones; nevertheless, our method can still achieve the leading SOTA performance on all tested benchmark datasets. Both the code and dataset are publicly available at https://github.com/wuzhenyubuaa/TSNet.**

***Index Terms*—Bi-stream fusion, image salient object detection, small-scale training set.**

I. Introduction

IMAGE salient object detection (ISOD) aims to well-segment the most attractive regions of the given image. As a pre-processing step, ISOD plays an important role in various computer vision tasks, such as visual tracking [1], [2], camouflaged object detection [3], [4], video saliency detection [5]–[8], and RGB-D completion [9]–[11].

Manuscript received July 28, 2020; revised October 29, 2020, December 7, 2020, and December 14, 2020; accepted December 14, 2020. Date of publication December 23, 2020; date of current version January 21, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 61802215, 61806106, 61672077, and 61532002, in part by the Natural Science Foundation of Shandong Province under Grants ZR2019BF011 and ZR2019QF009, in part by the National Science Foundation of the USA under Grants IIS-1715985 and IIS-1812606, and in part by the National Key R&D Program of China under Grant 2017YF-F0106407. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin. *(Corresponding author: Chenglizhao Chen.)*

Zhenyu Wu is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: wuzhenyu_961@126.com).

Shuai Li and Aimin Hao are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with Peng Cheng Laboratory (e-mail: lishuaiouc@126.com; ham_buaa@163.com).

Chenglizhao Chen is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China, and also with the Qingdao University, Shandong 266071, China (e-mail: cclz123@163.com).

Hong Qin is with the Stony Brook University, Stony Brook, NY 11794 USA (e-mail: qin@cs.stonybrook.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3046871.

Digital Object Identifier 10.1109/TMM.2020.3046871

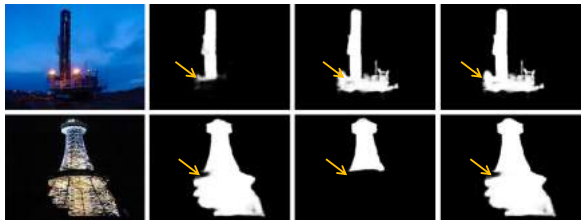

Fig. 1. Saliency maps generated by different network architectures might be complementary occasionally (see Table I), in which these saliency maps are obtained from the last convolutional layer of either VGG16 or ResNet50. The "Fusion" column shows the results obtained by fusing these two different backbones via the proposed gate control unit.

Inspired by cognitive psychology and neuroscience, the classical ISOD models [12], [13] are developed on fusing various hand-crafted saliency cues, however, all these cues fail to capture the wide variety of salient objects. After entering the deep learning era, the SOTA (**s**tate-**o**f-**t**he-**a**rt) ISOD performance has achieved tremendous improvement, which is mainly brought by both exquisitely crafted fancy network architectures [14]–[16] and newly available of large-scale well-annotated training sets [17], [18].

Following the single-stream network structure, recent ISOD methods [15], [16], [19]–[21] focused on how to effectively aggregate multi-level visual feature maps to boost their performances. Though remarkable progress has been achieved, these models might have reached to their performance limits, because they usually consist of a single feature backbone with limited ability in providing semantical information. Empirically, even for an identical image, different network architectures tend to have different feature responses. Inspired by this, we may easily achieve complementary semantical features if we simultaneously use two distinct feature backbones, where some pictorial demonstrations can be seen in Fig. 1.

In terms of the training dataset, the ISOD community has reached a consensus on the training protocol, i.e., models should be trained on the MSRA10K [17] or DUTS-TR [18] dataset, then tested on other datasets. However, we may raise a question regarding this widely-used training protocol, is this training strategy the best choice? According to our experimental results, some inspiring observations can be summarized as follows: 1) models' performances are not always positively correlated with the training data size, see the empirical results in Fig. 3; 2 the widely-used training sets (MSRA10 K and DUTS-TR) are also complementary with each other in performance, see quantitative

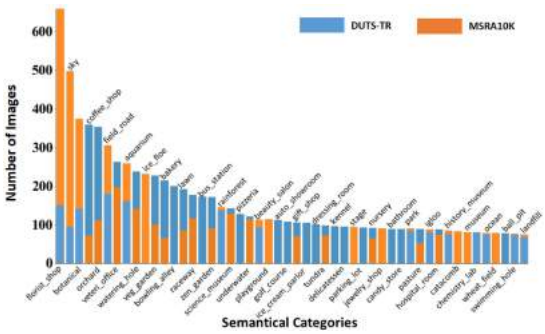

Fig. 2. Semantical category distribution (classified by [25]) of the MSRA10 K and the DUTS-TR sets. We only demonstrate the top-50 categories due to the limitation of space.

TABLE I
COMPARISONS OF THE THREE MOST REPRESENTATIVE SOTA MODELS TRAINED ON DIFFERENT DATASETS (AVERAGE F-MEASURE). WE USE **BOLD** TO EMPHASIZE THE BEST RESULTS

Trained on \ Tested on	Method	DUT-OMRON	DUTS-TE	ECSSD	HKU-IS	PASCAL-S
MSRA10K	PoolNet19 [22]	0.702	0.726	**0.888**	0.884	0.823
DUTS-TR		**0.738**	**0.781**	0.880	**0.895**	**0.846**
MSRA10K	CPD19 [23]	0.716	0.732	**0.882**	0.890	0.832
DUTS-TR		**0.738**	**0.784**	0.880	**0.903**	**0.836**
MSRA10K	AFNet19 [24]	**0.734**	**0.786**	**0.877**	**0.868**	0.752
DUTS-TR		0.729	0.772	0.871	0.856	**0.766**

results in Table I; 3) moreover, the MSRA10 K and DUTS-TR datasets are complementary in semantics as shown in Fig. 2.

Inspired by the aforementioned observations, this article constructs a novel small-scale training set named MD4K with total 4172 images, where all training instances are selected from either MSRA10 K or DUTS-TR and balanced in semantic category. Also, to take full advantage of this novel set, we devise a bi-stream network, where two different backbones (e.g., VGG16+ResNet50) are respectively used as the sub-branches. The behind rationale is to explore complementary semantical information which is already embedded in the pre-trained feature backbones for improving the SOTA performance.

To this end, we devise a novel gate control unit to effectively fuse complementary information encoded in different sub-branches. Meanwhile, we introduce a novel multi-layer attention into the bi-stream network for retaining tiny saliency details. We have also conducted extensive comparisons and component evaluations to show the advantages and effectiveness of the proposed approach (small-scale dataset & bi-stream network).

In summary, the contributions of this article can be summarized as follows:

- We provide a deeper insight into the interrelationship between ISOD performance and training data;
- We construct a new training set—being small-scale yet highly competitive in training performance;
- To take full advantage of the proposed small-scale training set (i.e., MD4K), we devise a novel fusion scheme for bi-stream network, in which the key technical components include the gate control unit and the multi-layer attention module;
- Extensive quantitative results demonstrate that the proposed model achieves the SOTA performance on all tested datasets, showing the effectiveness and superiority of the proposed method.

II. RELATED WORKS

To simulate the human visual attention, early ISOD methods mainly focus on designing various visual hand-crafted features, cues and priors, background cues, regional contrasts and other kinds of relevant low-level visual cues [26], [27]. Due to the space limitation, we only concentrate on the deep learning based ISOD models here. See [28], [29] for more details about traditional and early deep methods.

A. Single-Stream Models

Recently, most of the existing salient object detection models focus on aggregating multi-level/multi-scale features extracted from CNNs and push the performance of ISOD to a new level. As one of the earliest works, Hou *et al.* [15] proposed a top-down model to integrate both high-level and low-level features from different layers, achieving outstanding performance. Liu and Han [30] devised a coarse-to-fine approach, which locates salient objects firstly, then performs saliency refinement hierarchically and progressively for tiny saliency details. Following this rationale, various feature aggregation schemes [16], [30]–[38] were proposed subsequently. In contrary to the [15] which only uses specific-level features, Zhang *et al.* [32] integrated multi-level feature maps into different resolutions to predict saliency maps, aiming for incorporating both high-level semantic information and low-level spatial details simultaneously. Similarly, Wang *et al.* [36] integrated both top-down and bottom-up saliency inferences in an iterative and cooperative manner. Zhao *et al.* [37] presented an edge guidance network to model the complementary information provided by a single network. As a bridge between ISOD and fixation prediction, Wang *et al.* [39] built a novel attentive saliency network learning to detect salient objects from fixations, which narrows the gap between salient object detection and fixation prediction.

Our method is clearly different from the above approaches in two aspects. First, all of the above-mentioned models take the pre-trained classification network (e.g. ResNet and VGG) as a fixed feature extractor, ignoring the contributions of different encoder blocks. In sharp contrast, our model introduces a bi-stream encoder, where complementary information encoded in different networks can be learned mutually. Second, these models integrate multi-level features without considering their consistency, while our model with the proposed gate control unit can control the data flows between inter/intra-layers. The most closely related work to ours is [35] which proposed a gate function to control the data flows between different layers. Different from the gate settings proposed in [35], the major highlight of our gate control unit is that it can enable full interactions between two different sub-networks. Additionally, our gate control unit, a plug-in, can well retain the non-linear learning abilities of each individual sub-branch (see Section IV-A).

B. Bi-Stream Models

In recent years, the bi-stream network structure has achieved much research attention due to its effectiveness in broad computer vision applications, including video action recognition [40], image recognition [41], [42], and the one we

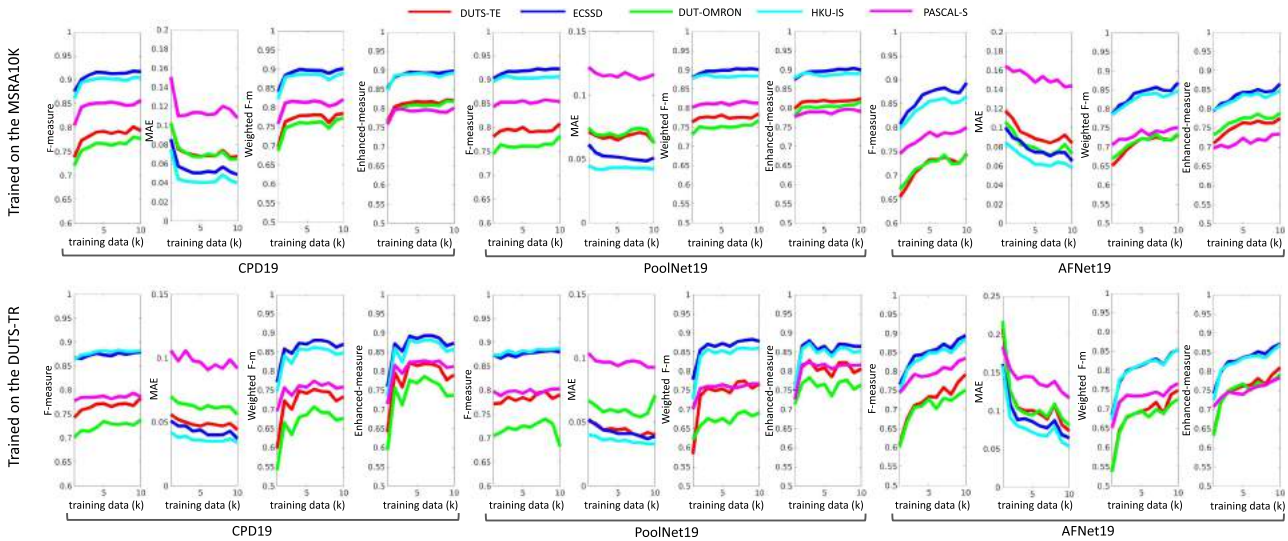

Fig. 3. Quantitative performances of three most representative SOTA models (CPD19 [23], PoolNet19 [22], and AFNet19 [24]) vary with the training data size (MSRA10 K and DUTS-TR), showing that the conventional consensus regarding the relationship between the model performance and the training set size—"*the model performance is positively related to the training set size*" may not always hold.

interested in, i.e., image salient object detection [43]–[46]. As a seminal work, Lin *et al.* [42] proposed a bilinear CNN model for the recognition task, which consists of two feature extractors to formulate image-level descriptor. Hou *et al.* [41] further presented a framework named DualNet to effectively learn more accurate representation for the image recognition task and its core idea is to coordinate two parallel DCNNs to learn complementary features. For human action recognition, Feichtenhofer *et al.* [40] proposed a bi-stream network with novel convolutional fusion layer between its sub-branches, aiming for incorporating both appearance and motion information.

Most recently, the bi-stream network has also been adopted in the ISOD community. Zhao *et al.* [43] proposed a multi-context deep learning framework, in which both the global and local contexts are combined in a unified deep learning framework. Zhang *et al.* [44] proposed a new deep neural network model named CapSal, which consists of two sub-networks to leverage the captioning information with both local and global visual contexts for predicting salient regions. In addition to using segmentation labels, researchers have also attempted to use the edge labels. For example, Su *et al.* [47] proposed a boundary-aware network to solve the selectivity-invariance dilemma of ISOD, where boundary localization and interior perception streams are introduced to capture features with selectivity and invariance, respectively. In [48], the work investigates the logical interrelations between binary segmentation and edge maps, which are then promoted to bidirectionally refine multi-level features of the two tasks. Similarly, Zhou *et al.* [45] proposed a lightweight two-stream model, in which one stream aims to learn the representations of salient regions and another focuses on the contours.

These approaches simply use the pre-trained network as a fixed feature extractor to extract common features, which are then processed by separate streams, and this topic has received less attention up to now. In sharp contrast, this article investigates the feature representation interrelations between different network structures, which aims to take advantage of complementary information presented in different networks to amend the probable failures that may occur in the indeterminate regions (see Fig. 1).

C. RGB-D ISOD Models

Previous works mainly focused on identifying salient regions via color channels (e.g., RGB) and achieved remarkable progress. However, the ISOD task is still challenging in some complex scenarios.

Recent literatures have shown that the depth information [9], [49]–[52] can be served as an important supplement to improve the ISOD performance, in which these works mainly relied on extracting salient features from RGB image and depth map separately, and then fused them in the shallow, middle, or deep layers of the network, and here we will list several most representative works. Piao *et al.* [53] introduced a novel depth-induced multi-scale recurrent attention network for saliency detection, which combined the RGB and depth complementary features in a multi-level fusion manner. Piao *et al.* [54] proposed a depth distiller, which explored the way of using network prediction and attention as two bridges to transfer depth knowledge from the depth stream to the RGB stream. Zhao *et al.* [55] proposed a unified framework for RGBD-based and RGB-based salient object detection tasks, which treated the depth information as supervision in the training stage. Zhang *et al.* [56] proposed a bilateral attention network to collaboratively learn complementary foreground and background features from both RGB and depth streams for better RGB-D ISOD performance. A detailed discussion of RGB-D based ISOD methods is beyond the main scope of this article, readers interested in RGB-D ISOD can refer to [57] for a comprehensive understanding.

D. Attention Mechanism

Inspired by the human visual system, attention mechanisms have been widely-used in various tasks such as object recognition [58], image captioning [59], visual question answering [60], pose estimation [61] and machine translation [62]. Xu *et al.* [59] introduced a soft deterministic and a hard stochastic attention mechanism for caption generation under a common framework. Xu *et al.* [60] proposed a novel multi-hop memory network with spatial attention for the VQA task which allows one to visualize the spatial inference process used by the deep network.

Due to attention mechanisms have a remarkable ability to select features, it is also suitable for saliency detection. As a pioneering work, Hu *et al.* [63] proposed the classic channel-wise attention to select the most representative feature channels. After that, this attention has been widely-applied in salient object detection. Recently, Zhang *et al.* [64] introduced both spatial-wise and channel-wise attention into the ISOD task. Wang *et al.* [29] devised an essential pyramid attention structure, which enables the network to concentrate more on salient regions when exploring multi-scale saliency information. Liu *et al.* [65] proposed a pixel-wise contextual attention mechanism to selectively combine global contexts with local ones. In [66], a novel reverse attention block was designed to highlight those image regions which were miss-detected before. Zhao *et al.* [67] proposed a novel saliency detection method, which contains a channel-wise attention module to capture context-aware multi-scale multi-receptive-field high-level features and a spatial attention module for low-level feature maps to refine salient object details.

However, those methods select features usually from low-level to high-level and ignore the relationship of high-level and low-level features. In sharp contrast to these works, the proposed multi-layer attention module transfers high-level semantic information to shallower layers to learn more detailed information, shrinking the given problem domain effectively. As a result, the proposed model learns more accurate details and achieves significant improvement.

E. Major Highlights of Our Method

In sharp contrast to the previous works which merely focus on the perspective of network design, our research might potentially be able to inspire the ISOD community to pay more attention on the training data aspect, which, in our view, could improve the SOTA performance more easily. Also, as another highlight, the proposed bi-stream network aims to take advantage of the rich semantic information embedded in the proposed small-scale MD4K set. To the best of our knowledge, this is the first time for a "wider" network trained on a small-scale dataset to outperform the existing modes trained on large-scale training sets.

III. A Small-Scale Training Set

Given an ISOD deep model, its performance usually relies on two factors: 1) the specific training dataset and 2) the corresponding set size. In fact, these two factors have been widely known [70], [71], while, in this article, we will provide some novel deeper insights.

A. Do We Really Need a Large-Scale Training Data?

The existing SOTA ISOD models usually have complex network architectures, thus these models heavily rely on large-scale training data to ensure their prominent performances. This issue motivates us to reconsider a basic problem regarding the ISOD task, i.e., will continually increasing the training data size be possible for achieving persistent performance improvements?

To clarify this issue, we have carried out a series of quantitative experiments on three SOTA ISOD models, including CPD19 [23], PoolNet19 [22] and AFNet19 [24]. We firstly train these models on the whole DUTS-TR(10 K)/MSRA10 K set, then retrain these models on smaller sets with 1000 images randomly removed each time, and this procedure will be repeated for multiple times. Thus, the relationship between the overall performance and the training data size can be observed in Fig. 3.

As we can see, when training data is increased to 2 K, the performances can be improved significantly. However, with the training data continue growing, the performance gains might become marginal, showing the fact that models' performances are not always positively correlated with the size of training data. We take the DUTS-TR training set for instance, the performance of CPD19 on the DUT-OMRON can be improved by about 12.5% after increasing the training data from 1 K to 2 K, while the performance gain decreases to 3.2% when increasing the training set size from 2 K to 3 K. Specifically, instead of using the entire DUTS-TR (10 K) set, the CPD19 achieves its best performance on the DUTS-TR (6 K).

Despite the aforementioned anomalies, the widely-used training sets (DUTS-TR and MSRA10 K) have two major limitations. First, the semantical category distributions of both training sets are unbalanced in essence. As we all known that a training set, which is balanced in its semantic distribution, is more preferable in producing better training performance. Actually, in the DUTS-TR set, there are 351 images classified into the "coffee shop" category, while, in sharp contrast, there are only 10 images can be classified into the "campus" category. Moreover, previous works [72], [73] have already demonstrated that the CNN based deep models are capable of understanding new concepts even only a few examples have been given, yet those redundant semantic scenes have less substantial help in improving the overall performance.

Second, there exists a large number of questionable binary annotations in the widely-used training sets, in which these annotations are easily leading to learning ambiguity. Thus, when constructing novel training set, we shall avoid including such annotations. Fig. 4 have summarized four types of questionable binary annotations. Concretely, the "Wrong Annotations" column refers to incorrectly labeling backgrounds as salient regions. The "Controversial Annotations" column illustrates that images containing no salient objects are mistakenly labeled with some possible salient regions. The "Conflict Annotations" column demonstrates the cases that salient regions are labeled following different labeling protocols. The "Grayscale Annotations" column shows the cases that salient regions are labeled with non-binary values which are positively related to the labeling confidences.

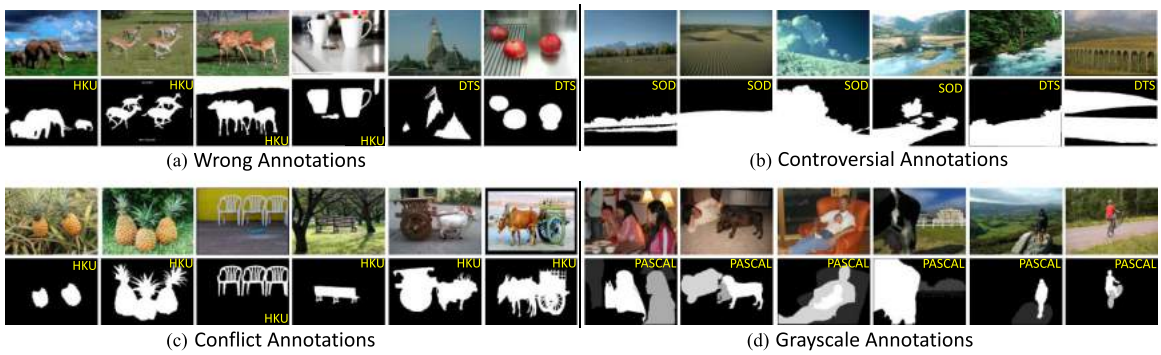

Fig. 4. Examples of inappropriate human annotations in ISOD benchmarks. The yellow marks in the top/bottom right of each image denote the corresponding dataset names, where {DTS, HKU, SOD, PASCAL} stands for {DUTS-TR [18], HKU-IS [43], SOD [68], and PASCAL-S [69]} sets respectively.

B. Which Training Set Should Be Selected?

In our ISOD community, SOTA models are usually trained on either MSRA10 K or DUTS-TR set in advance and tested later on others. However, this widely-used training/testing protocol suffers from a serious limitation; i.e., the inconsistent data distributions between different sets might result in the "domain-shift" problem.

For example, the images of the widely-used training set—MSRA10 K, are characterized with high contrast, center-surround, and simple background, and, in most cases, each image only contains a single salient object. However, the images in the widely-used testing set—PASCAL, are attributed as low contrast with complex background, in which multiple salient objects cases are occasionally existed in these images. Therefore, because of the inconsistencies mentioned above, models trained on MSRA10 K set usually perform worse on the PASCAL set, see the first row in Fig. 3. To conquer the "domain-shift" problem, we shall combine different training sets when constructing new training set, because, as we have mentioned in the introduction section, the widely-used training sets (MSRA10 K & DUTS-TR) are complementary in essence.

Specifically, previous works [18], [44], [74]–[76] have already demonstrated that semantic information, especially in cluttered scenes, is beneficial to the ISOD task. Meanwhile, it is also well known that a training set with good category distribution can ensure the given deep model to retain semantic-aware when striving for its saliency objective. Therefore, the balance of semantic categories is another key aspect that we need to take care when constructing new training set.

C. Our Novel Training Set (MD4K)

In this section, we build a small, GT bias-free and semantic category balanced training set, named MD4K, in which all training instances are selected from either MSRA10 K or DUTS-TR set.

We divide these two sets into 267 semantic categories via the off-the-shelf scene classification tool [25]. Then, we filter the dirty data, which consists of two parts: 1) 2 254 images that have been misclassified by the off-the-shelf scene classification tool [25], and 2) 72 images with biased annotations. Thus there are 9 012 left in the MSRA10 K set and 9 215 images left in the DUTS-TR set. As the major part, all those 2 254 images are filtered in a full automatical way. To be more specific, we resort to an auxiliary classifier with an identical structure to the primary classifier of the scene classification tool [25]. From the view of agreement maximization principle [79], [80], different classifiers would exhibit strong consistency usually, but such strong consistency tends to be vanished when facing some misclassified instances. Thus, we utilize the KL-divergence between the predictions of these two classifiers to show the classification confidence, where we automatically filter all those instances with classification confidences smaller than a relatively slack hard-threshold (0.67). As for the biased annotations, the only choice to filter these images might be the manual way, while, according to our quantitative evaluation result, removing this part of dirty data can only improve the overall performance slightly (about 0.2%).

We have noticed that the semantic category distribution of the images obeys the Pareto principle—20% scene categories account for 80% of the total. Specifically, the top-50 scene categories of MSRA10 K account for 71.23% of the whole MSRA10 K set, and such percentage is 74.13% in the DUTS-TR set. To balance the semantic categories, we randomly select a maximum of 40 images for each of the top-50 scene categories and then choose a maximum of 20 images for each of the remaining 217 scene categories. The main reason that we choose two different quantities (40/20) can be explained as follows and the corresponding experiments can be found in Table XI.

It is well known an ISOD model could achieve some performance gain if the category distribution of the training set is more consistent with that of the testing set. Since the existing testing sets also follow the Pareto principle, the ISOD models trained on datasets biased towards the top-50 scenes might perform better. Thus, when constructing the new MD4K dataset, we choose to use different quantities to ensure the semantic distribution of MD4K being similar to the original one, so that our MD4K set will have similar distribution to that of the testing sets.

In this way, we finally obtain a small-scale training set, containing 4 172 images with total of 267 semantical categories. The reason we choose 4 172 images is that we attempt to find a balance between training size and performance, and the performance trained on a different number of data is shown in Table II.

TABLE II
PERFORMANCE OF THE PROPOSED BI-STREAM MODEL TRAINED ON DIFFERENT SIZES OF THE MD4K SET

Tested on \ Trained on	DUTS-TR	MD1K	MD2K	MD3K	MD4K	MD5K	MD6K
DUT-OMRON [77]	0.835	0.715	0.794	0.832	0.857	0.864	0.866
DUTS-TE [18]	0.879	0.774	0.829	0.863	0.884	0.893	0.897
ECSSD [78]	0.934	0.876	0.876	0.918	0.945	0.947	0.955
HKU-IS [43]	0.933	0.864	0.885	0.920	0.942	0.948	0.952
PASCAL-S [69]	0.885	0.778	0.837	0.864	0.886	0.895	0.897

Though better performance can be achieved by using more training data, the overall performance improvements may gradually become really marginal (less than 0.4%). Moreover, by limiting the proposed set to a small size (4 K), it may be easier for future works to achieve further performance gain by adding other semantic-balanced data.

IV. PROPOSED NETWORK

In the previous sections, we have built a small-scale and semantic category balanced training set (MD4K), where this new set is capable of improving the SOTA performance occasionally (Table IX). To further improve, we propose a novel bi-stream network consisting of two different backbones, where these two feature backbones aim for providing complementary semantical information while taking full advantage of our MD4K set.

A. How to Fuse Bi-Stream Networks

In this section, we consider how to effectively fuse two different feature backbones. Our key rationale is to use feature maps extracted from one sub-branch to benefit another one. To facilitate a better understanding, we shall provide some preliminaries regarding the conventional fusion schemes in advance.

For simplicity, the function *f*: $\{(\mathbf{X}^R, \mathbf{X}^V) \rightarrow \mathbf{Y}\}$ represents fusing two feature maps $\mathbf{X}^R$ and $\mathbf{X}^V$ to generate the output feature $\mathbf{Y}$, where $\mathbf{X}^R$ and $\mathbf{X}^V$ respectively represent feature maps obtained from ResNet50 backbone and VGG16 backbone, $\{\mathbf{X}^R, \mathbf{X}^V, \mathbf{Y} \in \mathbb{R}^{H\times W\times C}\}$, H, W, C denote the height, width and channel respectively.

1) Element-Wise Summation: $\mathbf{Y}_{\rm sum}$, which calculates the sum of two features at the same location (w, h) and channel (c):

$$\mathbf{Y}_{\rm sum} = \sum_{c=1}^{C}\sum_{w=1}^{W}\sum_{h=1}^{H}(\mathbf{X}^R_{h,w,c} + \mathbf{X}^V_{h,w,c}). \tag{1}$$

2) Element-Wise Maximum: $\mathbf{Y}_{\rm max}$, which, analogously, computes the maximum of two input feature maps:

$$\mathbf{Y}_{\rm max} = \sum_{c=1}^{C}\sum_{w=1}^{W}\sum_{h=1}^{H} max(\mathbf{X}^R_{h,w,c}, \mathbf{X}^V_{h,w,c}). \tag{2}$$

3) Concatenation: $\mathbf{Y}_{\rm concat}$, which stacks the input feature maps channel-wisely:

$$\mathbf{Y}_{\rm concat} = Concat(\mathbf{X}^R_{h,w,c}, \mathbf{X}^V_{h,w,c}). \tag{3}$$

4) Convolution: $\mathbf{Y}_{conv}$, which first employs the concatenation operation to obtain features $\mathbf{Y}_{\rm concat} \in \mathbb{R}^{H\times W\times 2C}$ and then convolves it:

$$\mathbf{Y}_{conv} = \mathbf{Y}_{\rm concat} * \mathbf{W} + \mathbf{b}, \tag{4}$$

where $*$ denotes the convolution operation, $\mathbf{W}$ represents the convolution filters, and $\mathbf{b}$ denotes the bias parameters.

B. Bi-Stream Fusion Via GCU (*G*ate *C*ontrol *U*nit)

Generally, all of the above-mentioned fusion operations directly fuse two input feature maps without considering the feature conflictions between them, and this less consideration easily results in suboptimal results.

Inspired by the classic LSTM [81], we propose a novel gate control unit (input & output gates) to dynamically control the fusion process, and Fig. 5 illustrates the overall network architecture. In our method, the proposed input gate plays a critical role in aggregating feature maps. Let $\mathbf{X}^V = \{\mathbf{X}^V_i, i = 1, \ldots, 5\}$ denotes the feature maps for each convolutional block in the pre-trained VGG16 feature backbone, and, similarly, $\mathbf{X}^R$ represents that of the pre-trained ResNet50 backbone.

In our input gate, we use the dynamic thresholding to suppress those less-trustworthy input features. For example, each side-output of VGG16 with a probability below the threshold will be suppressed, where these side-outputs can be obtained via linear projections: $\{\mathbf{X}^V_i * \mathbf{W} + \mathbf{b}\}$, modulated by gates based on activation function (σ: sigmoid) as: $\{\sigma(\mathbf{X}^V_i * \mathbf{V}_{in} + \mathbf{b}_{in})\}$.

In practice, the input gate will be element-wisely multiplied by the side-output feature matrix, controlling the interactions between the parallel sub-branches hierarchically. Thus, the fused bi-stream feature maps ($\mathbf{Y}_{conv}$) can be obtained by using the below operation.

$$\begin{aligned}\Theta(\mathbf{X}^V_i) &= (\mathbf{X}^V_i * \mathbf{W} + \mathbf{b}) \otimes \sigma(\mathbf{X}^V_i * \mathbf{V}_{in} + \mathbf{b}_{in}),\\ \mathbf{Y}_{conv} &= f\left(\mathbf{X}^R_i, \Theta(\mathbf{X}^V_i)\right),\end{aligned} \tag{5}$$

where $\mathbf{W}$, $\mathbf{b}$, $\mathbf{V}_{in}$, $\mathbf{b}_{in}$ are learned parameters, σ is the sigmoid function and $\otimes$ is the element-wise multiplication operation.

Moreover, previous ISOD models directly propagate the feature maps from low-level layers to high-level layers without considering whether these features are beneficial to the ISOD task. In fact, only a small part of these features are useful, yet others may lead the fused performance even worse. To solve this problem, we propose a multiplicative operation based "output gate" to suppress those distractions from the non-salient regions. That is, given two consecutive layers, the feature responses in high-level layers $\sigma(\mathbf{X}^R_i * \mathbf{V}_{\rm out} + \mathbf{b}_{\rm out})$ will be served as the guidance for low-level layers $\mathbf{X}^R_{i-1} (i \in \{2, 3, 4, 5\})$ to adaptively determine which data flow should be propagated, and this procedure can be formulated as Eq. 6.

$$\tau(\mathbf{X}^R_i, \mathbf{X}^R_{i-1}) = \mathbf{X}^R_{i-1} \otimes \sigma(\mathbf{X}^R_i * \mathbf{V}_{\rm out} + \mathbf{b}_{\rm out}), \tag{6}$$

where $\mathbf{V}_{\rm out}$ and $\mathbf{b}_{\rm out}$ respectively represent the learned weights and biases. In this way, the salient regions with high feature responses can be enhanced, while the non-salient backgrounds can be suppressed in subsequent layers. In a word, our gate control unit is capable of boosting the conventional fusion

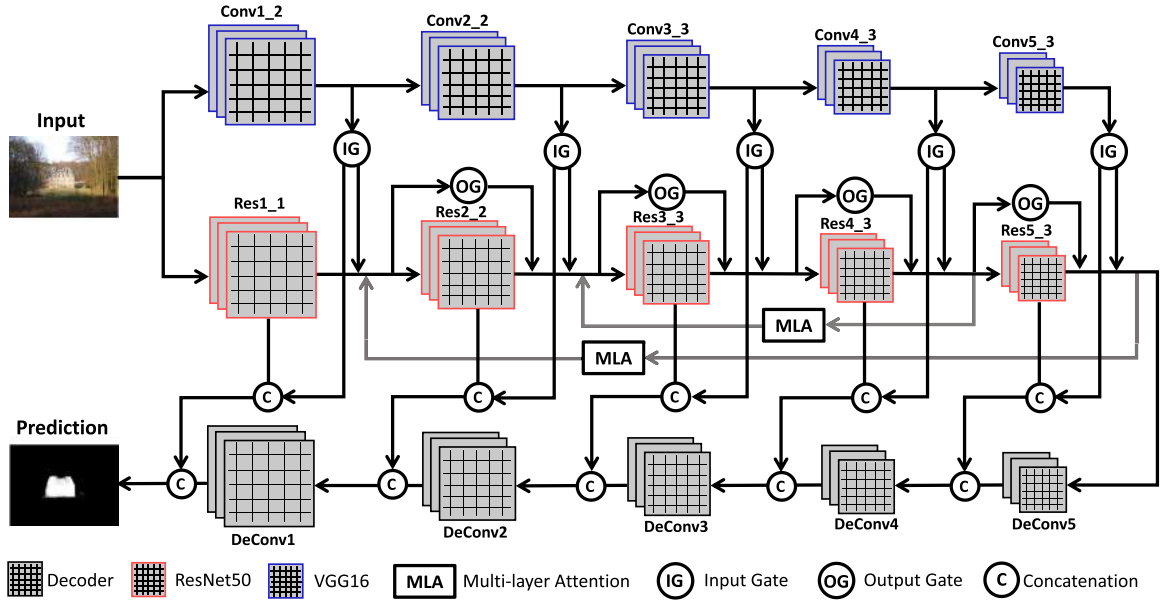

Fig. 5. Architecture of the proposed bi-stream network. Our bi-stream network is developed on the commonly used ResNet50 and VGG16, using both the newly designed gate control unit (Section. IV-A) and the scaling-free multi-layer attention (Section IV-D) to achieve the complementary status between two parallel sub-branches.

performances, and the quantitative evidences can be found in Section V.

C. Differences Between the Proposed GCU and the Gate Logic Used in LSTM

The error gradient in the LSTM [82] can be expressed as:

$$\nabla\left(tanh(\mathbf{X})\otimes\sigma(\mathbf{X})\right)=\sigma'(\mathbf{X})\nabla\mathbf{X}\otimes tanh(\mathbf{X}) + tanh'(\mathbf{X})\nabla\mathbf{X}\otimes\sigma(\mathbf{X}). \quad (7)$$

Notice that such gradient will gradually get vanished due to the down-scaling factor $tanh'(\mathbf{X})$ and $\sigma'(\mathbf{X})$. In sharp contrast, the gradient of our gate mechanism has a directional path $\nabla\mathbf{X}\otimes\sigma(\mathbf{X})$ without using any down-scaling operations for the activated gating units in $\sigma(\mathbf{X})$ as Eq. 8.

$$\nabla\left(\sigma(\mathbf{X})\otimes\mathbf{X}\right)=\nabla\mathbf{X}\otimes\sigma(\mathbf{X})+\sigma'(\mathbf{X})\nabla\mathbf{X}\otimes\mathbf{X}, \quad (8)$$

Thus, the proposed gate control unit outperforms the LSTM significantly (quantitative evidences can be found in Section V).

*D. The Proposed MLA (**M**ulti-**L**ayer **A**ttention)*

Generally, the predicted saliency maps tend to lose their details if we use sequential scaling operations (e.g., pooling). As we have mentioned before, visual features generated by deep layers tend to be dominated by high-level semantic information, while the shallower layers preserve low-level tiny details. Thus, the previous works have focused on devising feasible ways (e.g., short connections [15]) for integrating multi-level/multi-scale features.

However, as for our bi-stream network, the overall performance is mainly ensured by the exact fusion scheme (i.e., GCU), while the performances of its sub-branches (i.e., plain VGG16 and ResNet50) are clearly worse than other single-stream SOTA models. Consequently, the performance of our bi-stream network might degenerate if we follow the conventional "low ← high" or "high ← low" fusion schemes simply, because those low-quality feature maps tend to lead the fused ones even worse. Thus, we devise a novel **m**ulti-**l**ayer **a**ttention (MLA) mechanism on the ResNet50 sub-branch, of which the key rationale is to make full use of those features obtained in deep layers. Compared with the conventional "high ← low" scheme, the proposed MLA is very sparse, where only the high-level localization information (i.e., $\mathbf{X}_j^R, j\in\{4,5\}$) is adopted to complement the shallower layers directly.

The dataflow of the proposed MLA can be seen in Fig. 5, and its technical details can be formulated as follows:

$$\boldsymbol{\alpha}_j(l')=\frac{e^{\boldsymbol{\beta}_j(l')}}{\sum_{l=1}^{H\times W}e^{\boldsymbol{\beta}_j(l)}},\quad \boldsymbol{\beta}_j=tanh(\mathbf{X}_j^R*\mathbf{W}+\boldsymbol{b}), \quad (9)$$

where $\boldsymbol{\beta}_j\in\mathbb{R}^{H\times W}$ integrates the information of all channels in $\mathbf{X}_j^R$, $\boldsymbol{\beta}_j(l')$ denotes the feature at location l', and $\boldsymbol{\alpha}_j$ is the location attention map. Next, these location attention maps are applied to enhance those features in low-level layers $\mathbf{X}_m^R(m\in\{1,2\})$ as below.

$$\mathbf{X}_j^R\leftarrow f\left(\mathbf{X}_j^R, D\left((\mathbf{X}_m^R*\mathbf{W}+\mathbf{b})\otimes\boldsymbol{\alpha}_j\right)\right), \quad (10)$$

where the function $f(\cdot)$ denotes the element-wise summation, $D(\cdot)$ stands for down-sampling operation. The newly updated X_j^R will be feeded into the decoder to enhance spatial details progressively. In summary, compared with the widely used multi-scale short-connections, the proposed MLA is more suit for our bi-stream network.

V. Experiments and Results

A. Datasets

We evaluate the performance of the proposed method on six commonly used benchmark datasets, including DUT-OMRON [77], DUTS-TE [18], ECSSD [78], HKU-IS [43] and PASCAL-S [69] and SOC [71].

DUT-OMRON contains 5168 high-quality images. Images of this dataset have one or more salient objects with complex backgrounds.

DUTS-TE has 5019 images with high-quality pixel-wise annotations, selecting from the currently largest ISOD benchmark DUTS.

ECSSD has 1000 natural images, which contain many semantically meaningful and complex structures. As an extension of the complex scene saliency dataset, ECSSD is obtained by aggregating the images from BSD [89] and PASCAL VOC [90].

HKU-IS contains 4447 images. Most of the images in this dataset have low contrast with more than one salient object.

PASCAL-S contains 850 natural images with several objects, which are carefully selected from the PASCAL VOC dataset with 20 object categories and complex scenes.

SOC is designed to reflect the real-world scenes in detail. SOC is the largest instance-level ISOD dataset and contains 6000 images from more than 80 common categories.

B. Evaluation Metrics

We adopt five widely-used metrics to evaluate our method, including the precision-recall (PR) curves, the F-measure curves, mean absolute error (MAE), weighted F-measure, S-measure and E-measure.

PR curves: Following the previous settings [17], [91], we utilize the standard PR curves to evaluate the performance of our model.

F-measure: The F-measure is a harmonic mean of average precision and average recall. We compute the F-measure as

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{11}$$

where we set β^2 to be 0.3 to weigh precision more than recall.

MAE: The MAE is calculated as the average pixel-wise absolute difference between the binary GT and the saliency map S as Eq. 12.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \Big| S(x,y) - GT(x,y) \Big|, \tag{12}$$

where W and H are width and height of the saliency map S, respectively.

Weighted F-measure: Weighted F-measure [92] define weighted precision, which is a measure of exactness, and weighted recall, which is a measure of completeness:

$$F_\beta^w = \frac{(1+\beta^2) \times \text{Precision}^w \times \text{Recall}^w}{\beta^2 \times \text{Precision}^w + \text{Recall}^w}. \tag{13}$$

S-measure: S-measure [93] simultaneously evaluates region-aware S_r and object-aware S_o structural similarity between the saliency map and ground truth. It can be written as follows: $S_m = \alpha \times S_o + (1-\alpha) \times S_r$, where α is set to 0.5.

Enhanced-measure: Enhanced-measure (E-measure) [94] combines local pixel values with the image-level mean value to jointly evaluate the similarity between the prediction and the ground truth.

C. Comparison With the SOTA Models

We compare our model with 17 SOTA models, including DSS17 [15], Amulet17 [32], UCF17 [88], SRM17 [31], R^{3}Net18 [87], RADF18 [16], PAGRN18 [64], DGRL18 [33], MWS19 [86], CPD19 [23], AFNet19 [24], PoolNet19 [22], BASNet19 [85], EGNet19 [37], R^{2}Net20 [83], MRNet20 [84] and RANet20 [66]. For all of these SOTA models, the saliency maps are either generated by the original codes with recommended parameters or provided by the authors. Our results are generated by our model without using any additional processing.

1) Quantitative Comparisons: As a commonly used quantitative evaluation metric, we first investigate our model using the PR curves. As shown in the first row of Fig. 7, our model can consistently outperform the SOTA models on all tested benchmark datasets. Specifically, the proposed model outperforms other competitors on DUT-OMRON set significantly. Meanwhile, our model is evaluated by F-measure curves (see the second row of Fig. 7), which also demonstrates the superiority of our model. Moreover, the detailed experimental results in terms of five metrics (i.e, max F-measure, MAE, weighted F-measure, S-measure, and Enhanced F-measure) are listed in Table III and Table IV. As can be seen from these tables, our model shows good performance and outperforms other SOTA approaches significantly. In particular, in terms of max F-measure, the performance is improved by 5.8% over the second-best method RANet20 [66] on DUT-OMRON dataset.

2) Qualitative Comparisons: We demonstrate the qualitative comparisons in Fig. 6. The proposed method is capable of detecting salient objects accurately and completely. It can also adapt to various scenarios, including occlusion (the 1st row), complex background (the 2nd row), small object (3 rd row), and low contrast (4th row). Moreover, our method can highlight salient objects with sharp object boundaries.

3) Running Time and Model Complexity Comparisons: Table VI shows the running time comparisons. This evaluation was conducted on a machine with i7-6700 CPU, GTX 1070 GPU, where our model achieves 23 FPS (**f**rames **p**er **s**econd). Furthermore, we have compared our model with three most representative SOTA models in model size, FLOPs, and parameter number (Table VII).

Despite using two feature backbones, our model complexity is slightly worse than CPD [23]. As shown in Table VII, previous works have mainly focused on their decoders' design. In sharp contrast, our model concentrates on the encoder part solely, yet it has achieved the best performance.

4) Attributes-based Performance on SOC Set: As shown in Table V, we have compared the proposed method with several representative SOTA methods. Though we have reported the RANet20 [66], R^{2}Net20 [83] and MRNet20 [84] in other

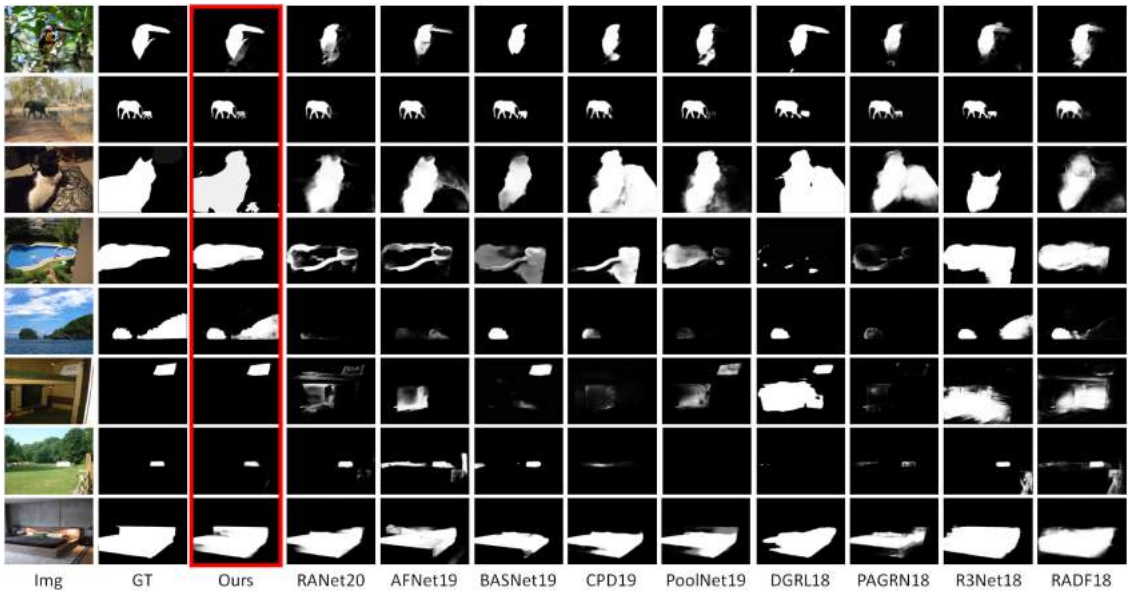

Fig. 6. Qualitative comparisons with the most recent SOTA models. Our approach can well locate salient objects accurately and completely with sharp object boundaries.

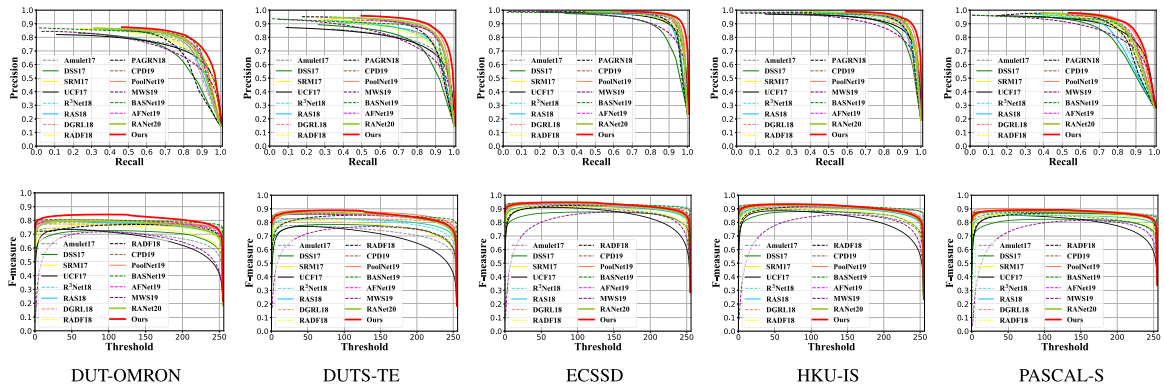

Fig. 7. Quantitative comparisons via PR curves (the first row) and F-measure curves (the second row).

datasets in the previous version, we have to omit these approaches in the SOC dataset, because these models are neither reported in the articles nor released with runnable codes currently. We can see that the proposed model outperforms almost all SOTA approaches significantly, demonstrating that the proposed model trained on MD4K can well adapt to various unseen categories.

D. Component Evaluations

1) Effectiveness of the Proposed MD4K Dataset: To illustrate the advantages of the proposed dataset, we present the evaluation results of the proposed models on our MD4K and DUTS-TR sets respectively in Table III and Table IV. Compared with using DUTS-TR as training set, our bi-stream network trained on MD4K set achieves the best performance in terms of different measures, showing the effectiveness of the proposed dataset. As shown in rows 9–14 of Table III, three SOTA models (i.e., PoolNet19, CPD19, and AFNet19) are trained on either the DUTS-TR dataset or our MD4K dataset. It can be observed that models trained on the MD4K dataset achieve better performances.

To demonstrate the effectiveness of balancing in semantic distribution, we also present the models' performances trained on MK4K and DTS4K, where images in these two sets are all **R**andom (R) or **C**ategory **B**alanced (CB) selected from either MSRA10 K or DUTS-TR. As shown in Table IX, models trained on the proposed semantic balanced MD4K achieves better performance than both MK4K and DTS4K sets. To be more specific, compared with the 'Random' sampling, applying the proposed 'Category Balanced' sampling scheme on single dataset (e.g., DTS4K† and MK4K†) could still improve the overall performance by 0.5%$\sim$1%. In fact, such limited improvement is mainly induced by the fact that the semantical category distribution in a single dataset is still very biased, even though the proposed sampling scheme has been used. In sharp contrast, as the main advocation of our article, the widely-used DUTS-TR and MSRA10 K training sets are complementary in essence, thus

TABLE III
DETAILED QUANTITATIVE COMPARISONS BETWEEN OUR METHOD AND 17 SOTA MODELS IN **F-MEASURE** AND **MAE** METRICS. TOP THREE SCORES ARE DENOTED IN RED, GREEN, AND BLUE, RESPECTIVELY. {MD4K, DTS, MK, MB, VOC, TH, CO} ARE TRAINING DATASETS WHICH RESPECTIVELY DENOTE {OUR SMALL DATASET, DUTS-TR, MSRA10 K, MSRA-B, PASCAL VOC2007, THUS10 K, AND MICROSOFT COCO}. THE SYMBOL "*" INDICATES THAT THE TARGET MODELS WERE TRAINED ON THE MD4K DATASET

Method	Backbone	Training		DUT-OMRON		DUTS-TE		ECSSD		HKU-IS		PASCAL-S	
		Images	Dataset	maxF_β ↑	MAE↓	maxF_β ↑	MAE↓	maxF_β ↑	MAE↓	maxF_β ↑	MAE↓	maxF_β ↑	MAE↓
Ours	ResNet50+VGG16	4172	MD4K	0.857	0.044	0.884	0.038	0.945	0.036	0.942	0.031	0.886	0.082
Ours	ResNet50+VGG16	10553	DTS	0.835	0.046	0.879	0.041	0.934	0.039	0.933	0.033	0.885	0.089
Ours	ResNet50+VGG16	10000	MK	0.828	0.047	0.863	0.044	0.931	0.042	0.917	0.035	0.857	0.088
Ours	ResNet50+ResNet50	4172	MD4K	0.833	0.046	0.855	0.041	0.921	0.043	0.916	0.037	0.853	0.087
Ours	VGG16+VGG16	4172	MD4K	0.826	0.049	0.849	0.047	0.924	0.042	0.918	0.033	0.844	0.092
RANet20 [66]	VGG16	10553	DTS	0.799	0.058	0.874	0.044	0.941	0.042	0.928	0.036	0.866	0.078
R^2Net20 [83]	VGG16	10553	DTS	0.793	0.061	0.855	0.050	0.935	0.044	0.921	0.030	0.864	0.075
MRNet20 [84]	ResNet50	10553	DTS	0.731	0.062	0.792	0.048	0.904	0.048	0.891	0.039	0.818	0.075
CPD19* [23]	ResNet50	4172	MD4K	0.762	0.052	0.850	0.040	0.934	0.037	0.915	0.032	0.846	0.090
CPD19 [23]	ResNet50	10553	DTS	0.754	0.056	0.841	0.044	0.926	0.037	0.911	0.034	0.843	0.092
PoolNet19* [22]	ResNet50	4172	MD4K	0.767	0.051	0.863	0.042	0.931	0.040	0.922	0.033	0.859	0.084
PoolNet19 [22]	ResNet50	10553	DTS	0.763	0.055	0.858	0.040	0.920	0.042	0.917	0.033	0.856	0.093
AFNet19* [24]	VGG16	4172	MD4K	0.765	0.054	0.842	0.044	0.932	0.041	0.913	0.034	0.854	0.087
AFNet19 [24]	VGG16	10553	DTS	0.759	0.057	0.838	0.046	0.924	0.042	0.910	0.036	0.852	0.089
BASNet19 [85]	ResNet34	10553	DTS	0.805	0.057	0.859	0.048	0.942	0.037	0.929	0.032	0.876	0.092
EGNet19 [37]	ResNet50	10053	DTS	0.815	0.053	0.888	0.040	0.943	0.037	0.935	0.031	0.869	0.090
MWS19 [86]	DenseNet169	310K	CO+DTS	0.677	0.109	0.722	0.092	0.859	0.096	0.835	0.084	0.781	0.153
PAGRN18 [64]	VGG19	10553	DTS	0.707	0.071	0.818	0.056	0.904	0.061	0.897	0.048	0.817	0.120
DGRL18 [33]	ResNet50	10553	DTS	0.739	0.062	0.806	0.051	0.914	0.049	0.900	0.036	0.856	0.085
RADF18 [16]	VGG16	10000	MK	0.756	0.072	0.786	0.072	0.905	0.060	0.895	0.050	0.817	0.123
R^3Net18 [87]	ResNeXt	10000	MK	0.460	0.138	0.478	0.136	0.656	0.161	0.583	0.150	0.611	0.203
SRM17 [31]	ResNet50	10553	DTS	0.725	0.069	0.799	0.059	0.905	0.054	0.893	0.046	0.812	0.105
Amulet17 [32]	VGG16	10000	MK	0.715	0.098	0.751	0.085	0.904	0.059	0.884	0.052	0.836	0.107
UCF17 [88]	VGG16	10000	MK	0.705	0.132	0.740	0.118	0.897	0.078	0.871	0.074	0.820	0.131
DSS17 [15]	VGG16	2500	MB	0.681	0.092	0.751	0.081	0.856	0.090	0.865	0.067	0.777	0.149

TABLE IV
CONTINUED QUANTITATIVE COMPARISONS IN TERMS OF **WEIGHTED F-MEASURE**, **S-MEASURE**, AND **E-MEASURE**

Method	Backbone	Training		DUT-OMRON			DUTS-TE			ECSSD			HKU-IS			PASCAL-S		
		Images	Dataset	W-F_β ↑	S-m↑	E-m	W-F_β ↑	S-m↑	E-m	W-F_β ↑	S-m↑	E-m	W-F_β ↑	S-m↑	E-m	W-F_β ↑	S-m↑	E-m
Ours	ResNet50+VGG16	4172	MD4K	0.761	0.858	0.809	0.804	0.883	0.854	0.915	0.936	0.917	0.902	0.921	0.912	0.816	0.857	0.843
Ours	ResNet50+VGG16	10553	DTS	0.757	0.847	0.803	0.788	0.871	0.847	0.908	0.920	0.911	0.893	0.914	0.906	0.808	0.851	0.845
Ours	ResNet50+VGG16	10000	MK	0.748	0.843	0.792	0.782	0.864	0.837	0.902	0.915	0.907	0.884	0.907	0.891	0.794	0.842	0.838
Ours	ResNet50+ResNet50	4172	MD4K	0.723	0.834	0.790	0.782	0.861	0.841	0.891	0.918	0.903	0.886	0.907	0.892	0.803	0.848	0.842
Ours	VGG16+VGG16	4172	MD4K	0.716	0.831	0.785	0.780	0.867	0.835	0.890	0.912	0.906	0.874	0.904	0.887	0.788	0.827	0.821
RANet20 [66]	VGG16	10553	DTS	0.671	0.825	0.742	0.743	0.874	0.776	0.866	0.917	0.844	0.846	0.908	0.841	0.757	0.847	0.812
R^2Net20 [83]	VGG16	10553	DTS	-	0.824	-	-	0.861	-	-	0.915	-	-	0.903	-	-	0.847	-
CPD19* [23]	ResNet50	4172	MD4K	0.722	0.845	0.793	0.785	0.874	0.844	0.891	0.913	0.905	0.879	0.912	0.894	0.784	0.839	0.835
CPD19 [23]	ResNet50	10553	DTS	0.705	0.825	0.787	0.769	0.868	0.838	0.889	0.918	0.902	0.866	0.906	0.888	0.771	0.828	0.827
PoolNet19* [22]	ResNet50	4172	MD4K	0.717	0.851	0.785	0.786	0.894	0.822	0.893	0.940	0.876	0.885	0.923	0.878	0.798	0.849	0.828
PoolNet19 [22]	ResNet50	10553	DTS	0.696	0.831	0.775	0.775	0.886	0.819	0.890	0.926	0.877	0.873	0.919	0.870	0.781	0.847	0.826
AFNet19* [24]	VGG16	4172	MD4K	0.712	0.834	0.764	0.762	0.874	0.788	0.875	0.916	0.853	0.863	0.912	0.844	0.787	0.845	0.816
AFNet19 [24]	VGG16	10553	DTS	0.690	0.826	0.760	0.747	0.866	0.785	0.867	0.914	0.849	0.848	0.905	0.839	0.772	0.833	0.810
BASNet19 [85]	ResNet34	10553	DTS	0.752	0.836	0.857	0.793	0.865	0.886	0.904	0.916	0.938	0.889	0.909	0.936	0.776	0.819	0.834
EGNet19 [37]	ResNet50	10053	DTS	0.701	0.841	0.760	0.769	0.886	0.802	0.887	0.925	0.870	0.870	0.918	0.860	0.777	0.835	0.821
MWS19 [86]	DenseNet169	310K	CO+DTS	0.423	0.756	0.336	0.531	0.757	0.610	0.652	0.828	0.555	0.613	0.818	0.508	0.613	0.753	0.546
PAGRN18 [64]	VGG19	10553	DTS	0.601	0.775	0.604	0.685	0.837	0.613	0.822	0.889	0.558	0.805	0.887	0.507	0.701	0.793	0.592
DGRL18 [33]	ResNet50	10553	DTS	0.709	0.806	0.843	0.768	0.841	886	0.891	0.903	0.937	0.875	0.895	0.938	0.791	0.828	0.838
RADF18 [16]	VGG16	10000	MK	0.611	0.813	0.603	0.635	0.824	0.619	0.802	0.895	0.717	0.782	0.888	0.707	0.709	0.797	0.732
R^3Net18 [87]	ResNeXt	10000	MK	0.726	0.817	0.840	0.648	0.835	0.700	0.902	0.910	0.942	0.877	0.895	0.935	0.737	0.788	0.802
SRM17 [31]	ResNet50	10553	DTS	0.607	0.798	0.677	0.662	0.835	0.711	0.825	0.895	0.813	0.802	0.888	0.799	0.736	0.817	0.776
Amulet17 [32]	VGG16	10000	MK	0.563	0.781	0.542	0.594	0.803	0.575	0.798	0.894	0.729	0.767	0.883	0.703	0.732	0.820	0.680
UCF17 [88]	VGG16	10000	MK	0.465	0.758	0.342	0.493	0.778	0.360	0.688	0.883	0.445	0.656	0.866	0.436	0.666	0.808	0.407
DSS17 [15]	VGG16	2500	MB	0.481	0.748	0.285	0.538	0.790	0.303	0.688	0.836	0.329	0.677	0.852	0.291	0.626	0.749	0.447

our MD4K set (MD4K†) constructed on these two sets could achieve significant performance improvement.

2) Effectiveness of the Proposed Bi-stream Network: To demonstrate the effectiveness of the proposed bi-stream network, we additionally implement the proposed bi-stream network by using two identical feature backbones, i.e., "VGG16+VGG16" and "ResNet50+ResNet50," see Table III. Compared with the "VGG16+VGG16" and "ResNet50+ResNet50" models, the proposed bi-stream network achieves better performance. Besides, we also report the performance of the proposed bi-stream network trained on the DUTS-TR dataset (the 2nd row of Table III), where our model achieves better performance than other SOTA models, showing the effectiveness of the proposed bi-stream network.

To further illustrate the complementarity between VGG16 and ResNet50, Fig. 8 have provided some qualitative demonstrations, in which the proposed bi-stream network is capable of revealing different but complementary salient regions.

TABLE V
ATTRIBUTES-BASED PERFORMANCE ON SOC DATASET [71]. WE USE THE S-MEASURE TO EVALUATE EACH SPECIFIC ATTRIBUTE AND THE AVERAGE PERFORMANCE IS GIVEN IN THE LAST ROW. TOP THREE SCORES ARE DENOTED IN RED, GREEN, AND BLUE, RESPECTIVELY

Attr	DSS17	Amulet17	SRM17	RAS18	R^3Net18	DGRL18	CPD19	PoolNet19	BASNet19	Ours
AC	0.744	0.756	0.794	0.694	0.703	0.791	0.801	0.791	0.804	0.803
BO	0.587	0.653	0.691	0.475	0.451	0.728	0.695	0.596	0.638	0.783
CL	0.689	0.718	0.747	0.619	0.680	0.756	0.768	0.755	0.742	0.804
HO	0.753	0.764	0.794	0.692	0.715	0.800	0.810	0.808	0.791	0.821
MB	0.758	0.756	0.817	0.691	0.696	0.827	0.854	0.819	0.818	0.823
OC	0.703	0.714	0.734	0.616	0.643	0.748	0.766	0.745	0.744	0.778
OV	0.702	0.744	0.775	0.622	0.639	0.778	0.785	0.765	0.774	0.815
SC	0.752	0.748	0.774	0.697	0.703	0.779	0.790	0.793	0.766	0.767
SO	0.707	0.675	0.727	0.678	0.686	0.727	0.753	0.760	0.726	0.734
Avg	0.719	0.715	0.757	0.664	0.683	0.759	0.780	0.759	0.756	0.808

TABLE VI
SPEED COMPARISONS, FPS: **F**RAMES **P**ER **S**ECOND

Method	Ours	RANet20	R^2Net20	MRNet20	BASNet19
FPS	**23**	42	33	14	25
Method	CPD19	PoolNet19	AFNet19	DGRL18	RADF18
FPS	62	27	23	6	18

TABLE VII
COMPARISONS IN MODEL SIZE, FLOPS AND HIDDEN PARAMETER SIZE

Method	Model(MB)	Encoder(MB)	Decoder(MB)	FLOPs(G)	Params(M)
Ours	235.5	152.6	82.9	65.53	71.67
CPD19 [23]	192	95.6	96.4	17.75	47.85
BASNet19 [85]	348.5	87.3	261.2	127.32	87.06
PoolNet19 [22]	278.5	94.7	183.8	88.91	68.26

TABLE VIII
PERFORMANCE COMPARISONS OF DIFFERENT FUSION STRATEGIES, WHERE "W/" DENOTES "WITH," "W/O" DENOTES "WITHOUT"; GCU: GATE CONTROL UNIT; CONV, SUM, CONCAT, MAX ARE FOUR CONVENTIONAL FUSION SCHEMES MENTIONED IN SECTION IV-A. "CONV W/ GCU (LSTM)" DENOTES THE PERFORMANCE USING THE GATE CONTROL LOGIC OF LSTM

Fusion Method	DUT-OMRON		DUTS-TE		ECSSD	
	maxF_β	MAE	maxF_β	MAE	maxF_β	MAE
Conv w/ GCU (Ours)	**0.857**	**0.044**	**0.884**	**0.038**	**0.945**	**0.036**
Conv w/ GCU (LSTM)	0.834	0.046	0.864	0.045	0.934	0.042
Conv w/o GCU	0.821	0.049	0.844	0.051	0.927	0.048
Sum w/ GCU	**0.848**	**0.047**	**0.873**	**0.044**	**0.925**	**0.043**
Sum w/o GCU	0.813	0.055	0.845	0.052	0.897	0.049
Concat w/ GCU	**0.827**	**0.049**	**0.862**	**0.047**	**0.908**	**0.046**
Concat w/o GCU	0.802	0.059	0.847	0.058	0.887	0.054
Max w/ GCU	**0.818**	**0.050**	**0.853**	**0.048**	**0.909**	**0.047**
Max w/o GCU	0.813	0.054	0.836	0.054	0.887	0.053

*3) Effectiveness of the GCU (**G**ate **C**ontrol **U**nit):* To validate the exact contribution of the proposed GCU, we take the above-mentioned fusion methods mentioned in Section IV-A as the baselines. Then, we apply the proposed GCU into these conventional fusion schemes, and the corresponding quantitative results are shown in Table VIII. It can be seen that these conventional fusion schemes equipped with GCU are clearly better than their plain versions.

TABLE IX
QUANTITATIVE EVALUATION REGARDING THE EFFECTIVENESS OF THE PROPOSED SMALL-SCALE TRAINING SET MD4K. DTS4K/MK4K REPRESENTS EXTRACTING 4172 IMAGES FROM DUTS-TR/MSRA10 K SET. THE SYMBOL '$*$' INDICATES **R**ANDOM (R) SAMPLING FROM THE SOURCE DATASET WHILE '†' STANDS FOR THE **C**ATEGORY **B**ALANCED (CB) SAMPLING SCHEME THAT IS IDENTICAL TO THE SAMPLING SCHEME USED IN CONSTRUCTING OUR MD4K SET

Method	Training		DUT-OMRON		DUTS-TE		ECSSD	
	Sampling	Dataset	max$F_\beta\uparrow$	MAE↓	max$F_\beta\uparrow$	MAE↓	max$F_\beta\uparrow$	MAE↓
Ours	CB	MD4K†	**0.857**	**0.044**	**0.884**	**0.038**	**0.945**	**0.036**
Ours	CB	DTS4K†	0.832	0.047	0.843	0.047	0.913	0.045
Ours	R	DTS4K*	0.825	0.048	0.838	0.051	0.905	0.048
Ours	CB	MK4K†	0.823	0.056	0.825	0.050	0.895	0.049
Ours	R	MK4K*	0.820	0.060	0.823	0.052	0.887	0.050
CPD19	CB	MD4K†	**0.762**	**0.052**	**0.850**	**0.040**	**0.934**	**0.037**
CPD19	CB	DTS4K†	0.733	0.061	0.832	0.046	0.910	0.041
CPD19	R	DTS4K*	0.721	0.063	0.824	0.048	0.902	0.043
CPD19	CB	MK4K†	0.731	0.062	0.824	0.052	0.896	0.054
CPD19	R	MK4K*	0.722	0.060	0.818	0.056	0.889	0.061
PoolNet19	CB	MD4K†	**0.767**	**0.051**	**0.863**	**0.042**	**0.931**	**0.040**
PoolNet19	CB	DTS4K†	0.743	0.061	0.845	0.045	0.912	0.040
PoolNet19	R	DTS4K*	0.738	0.064	0.839	0.047	0.907	0.043
PoolNet19	CB	MK4K†	0.738	0.063	0.838	0.048	0.903	0.044
PoolNet19	R	MK4K*	0.733	0.065	0.836	0.048	0.897	0.045
AFNet19	CB	MD4K†	**0.765**	**0.054**	**0.842**	**0.044**	**0.932**	**0.041**
AFNet19	CB	DTS4K†	0.745	0.062	0.831	0.053	0.908	0.051
AFNet19	R	DTS4K*	0.737	0.065	0.823	0.057	0.891	0.062
AFNet19	CB	MK4K†	0.732	0.058	0.834	0.049	0.910	0.050
AFNet19	R	MK4K*	0.728	0.063	0.830	0.053	0.895	0.060

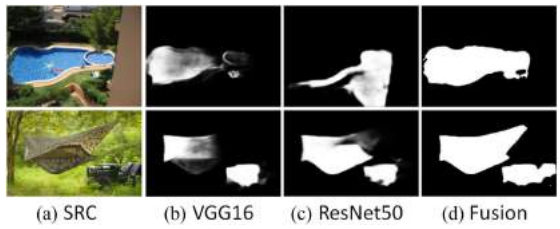

Fig. 8. Qualitative demonstrations to show the ability of the proposed bi-stream network in achieving complementary fusion status between its VGG16 and ResNet50 sub-branches.

*4) Effectiveness of the MLA (**M**ulti-**l**ayer **A**ttention):* As shown in the last row of Table IX, the proposed MLA improves the overall performance significantly. In particular, in terms of F-measure and MAE, the performance on DUT-OMRON set is improved by 2.3% and 6% respectively. Additionally, Fig. 9 shows that the proposed MLA is capable of sharping object boundaries.

5) Why Do We Choose the ResNet50 as the Main Sub-branch?: As shown in the first row of Table X, we have carried out the experiments of $ResNet50 \rightarrow VGG16$, which means feeding

TABLE X
QUANTITATIVE EVALUATIONS ON TWO DIFFERENT FUSION SCHEMES. $VGG16 \rightarrow ResNet50$ DENOTES THAT FEATURES OF VGG16 ARE FEEDED INTO THE RESNET50, OR VICE VERSA. THE BEST RESULT IS SHOWN IN **BOLD**

Method	DUT-OMRON		DUTS-TE		ECSSD		HKU-IS		PASCAL-S	
	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$
VGG16 $\rightarrow$ ResNet50 (Ours)	**0.857**	**0.044**	**0.884**	**0.038**	**0.945**	**0.036**	**0.942**	**0.031**	0.886	**0.082**
ResNet50 $\rightarrow$ VGG16	0.843	0.045	0.876	0.042	0.937	0.038	0.935	0.034	**0.889**	0.085
Shallow Fusion (Ours)	**0.857**	**0.044**	**0.884**	**0.038**	**0.945**	**0.036**	**0.942**	**0.031**	**0.886**	**0.082**
Deep Fusion	0.839	0.047	0.862	0.043	0.928	0.040	0.931	0.035	0.871	0.092

TABLE XI
QUANTITATIVE EVALUATIONS OF DIFFERENT CATEGORY COMBINATION RATIOS. GS: GAUSSIAN SAMPLING

NO.	Ratio	DUT-OMRON		DUTS-TE		ECSSD		HKU-IS		PASCAL-S	
		max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$	max$F_\beta \uparrow$	MAE$\downarrow$
Model-1	20:20	0.842	0.047	0.870	0.045	0.936	0.043	0.935	0.034	0.868	0.086
Model-2	40:20	0.857	0.044	0.884	0.038	0.945	0.036	0.942	0.031	0.886	0.082
Model-3	60:20	0.858	0.043	0.892	0.037	0.947	0.035	0.947	0.031	0.895	0.081
Model-4	40:10	0.836	0.047	0.872	0.043	0.938	0.042	0.931	0.033	0.874	0.085
Model-5	GS	0.825	0.055	0.865	0.047	0.932	0.044	0.924	0.037	0.859	0.090

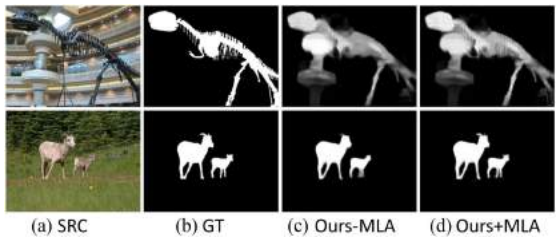

Fig. 9. Visual comparison of the proposed model with multi-layer attention ("Ours+MLA") and without multi-layer attention ("Ours-MLA").

the ResNet50's multi-scale features into that of the VGG16, and we have observed a clear performance degeneration after switching the roles of the VGG16 and ResNet50 sub-branch. We noticed that an ISOD model taking ResNet50 as backbone usually outperforms the VGG16 based version. Thus, it is quite normal for the proposed bi-stream network to be degenerated after switching its main backbone from ResNet50 to VGG16, because the main feature extractor (i.e., the subbranch who receives complementary information) is the key factor influencing the overall performance.

6) Shallow Fusion vs. Deep Fusion: We also implement the deep fusion version of the proposed bi-stream network, where fusion processed are mainly performed in the decoder part. For a fair comparison, the deep fusion network have also adopted both GCU and MLA, which is completely identical to that of the proposed shallow fusion version (fused in the encoder part). As shown in the last row of Table X, the shallow fusion version outperforms the deep fusion version persistently for all cases. This quantitative result further confirms the superiority of the proposed model in extracting and fusing multi-level paired complementary information.

7) The Effects of Different Ratio of the Proposed MD4K: We also report how the performance will be impacted if a different ratio of the distribution is used. The corresponding quantitative results can be found in Table XI, where several sampling ratios: {20:20}, {40:20}, {60:20}, {40:10}, and the Gaussian scheme have been tested. Taking the ratio {40:20} for instance, in which we select a maximum of 40 images for each of the top-50 scene category, while, for each of the remaining 217 scene category, a maximum of 20 images will be included.

As can be seen in Table XI, the performance tends to improve if we adopt a ratio biasing towards the top-50 scene category, but, with the increasing of the biasing tendency, the performance gap would become very marginal. For example, compared with the bias-free scheme (i.e., {20:20}, Model-1), we can easily notice that the Model-2 ({40:20}) outperforms it significantly. Moreover, the very slight performance gain achieved by the {60:20} ratio is mainly induced by the performance trade-off between the increasing of training instances and the over-biased semantic distribution. In addition, to verify the effectiveness of sampling instances from the 217 minor categories, we have tested the ratio {40:10}, where we have observed a significant performance decrease.

Further, we have also tested the widely-used Box − Muller Gaussian to estimate the category distribution, and then sampling our small-scale MD4K according to this distribution, where the quantitative results can be seen in the bottom row of Table XI. As can be seen, the 'Gaussian Sampling' could degenerate the overall performance by a large margin, and the main reason is that the imbalanced category distribution.

In a word, we recommend the ratio {40:20} for the proposed MD4K set to find a balance between training set size, semantic category distribution, and performance.

E. Failure Case and Analysis

We have provided some failure cases in Fig. 10. Compared with the conventional single-stream approaches, the major advantage of the proposed bi-stream network is its capability of taking full complementary fusion between the parallel subbranches. Since subbranches with different feature backbones can complement with each other, the fused saliency maps outperforms either of them easily. However, one major limitation still exists, i.e., our approach may produce failure detections

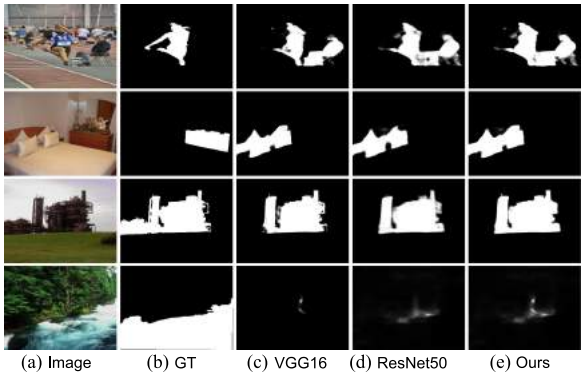

Fig. 10. Failure cases of our proposed bi-stream network.

when both of its subbranches have failed in providing correct saliency cues.

VI. Conclusion

In this article, we have provided a deeper insight into the interrelationship between the ISOD performance and the training dataset. Inspired by our observations, we build a small, hybrid, and semantic category balanced new training set. This new set is able to improve the SOTA performances extensively, providing a paradigm regarding how to effectively design a training set for performance gain. Meanwhile, we have proposed a novel bi-stream architecture with gate control unit and multi-layer attention to take full advantage of the proposed small-scale training set. Extensive quantitative comparisons and component evaluations have demonstrated that the proposed bi-stream network trained on the new small-scale training set can achieve new SOTA performance.

References

[1] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognit.*, vol. 48, no. 9, pp. 2885–2905, 2015.

[2] C. Chen, S. Li, and H. Qin, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognit.*, vol. 52, pp. 410–432, 2016.

[3] D.-P. Fan *et al.*, "Camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3052–3062.

[4] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 263–273.

[5] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "A plug-and-play scheme to adapt image saliency deep model for video data," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2020.3023080.

[6] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial-temporal information," *IEEE Trans. Image Process.*, vol. 29, pp. 1090–1100, 2019.

[7] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[8] Y. Li, S. Li, C. Chen, H. Qin, and A. Hao, "Accurate and robust video saliency detection via self-paced diffusion," *IEEE Trans. Multimedia*, vol. 22, pp. 1153–1167, 2019.

[9] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4296–4307, 2020.

[10] X. Wang *et al.*, "Data-level recombination and lightweight fusionscheme for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 458–471, 2021.

[11] X. Wang, S. Li, C. Chen, A. Hao, and H. Qin, "Knowing depth quality in advance: A depth quality assessment method for RGB-D salient object detection," *Neurocomputing*, to be published, doi: j.neucom.2020.12.071.

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

[13] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2303–2316, Aug. 2015.

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

[15] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5300–5309.

[16] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. Assoc. Adv. Artif. Intell.*, 2018, pp. 6943–6950.

[17] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

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

[19] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 324–336, Feb. 2020.

[20] K. Fu, Q. Zhao, I. Y. Gu, and J. Yang, "Deepside: A general deep framework for salient object detection," *Neurocomputing*, vol. 356, pp. 69–82, 2019.

[21] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9413–9422.

[22] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3917–3926.

[23] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3907–3916.

[24] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1623–1632.

[25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A. 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[26] C. Deng, X. Yang, F. Nie, and D. Tao, "Saliency detection via a multiple self-weighted graph-based manifold ranking," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 885–896, Apr. 2020.

[27] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 750–762, Apr. 2017.

[28] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient Object detection: A survey," *Comput. Vis. Media*, vol. 1411, pp. 1–34, 2019.

[29] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1448–1457.

[30] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 678–686.

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

[32] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

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

[34] L. Ye *et al.*, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, Aug. 2017.

[35] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1741–1750.

[36] W. Wang, J. Shen, M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5968–5977.

[37] J. Zhao *et al.*, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8779–8788.

[38] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 247–256.

[39] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

[40] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[41] S. Hou, X. Liu, and Z. Wang, "DualNet: Learn complementary features for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 502–510.

[42] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

[43] 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.

[44] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "CapSal: Leveraging captioning to boost semantics for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6024–6033.

[45] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1449–1457.

[46] Z. Wu, S. Li, C. Chen, A. Hao, and H. Qin, "Recursive multi-model complementary deep fusion forrobust salient object detection via parallel sub networks," 2020, *arXiv:2008.04158*.

[47] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.

[48] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7264–7273.

[49] J.-X. Zhao *et al.*, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3927–3936.

[50] J. Zhang *et al.*, "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8582–8591.

[51] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3052–3062.

[52] X. Wang *et al.*, "Data-level recombination and lightweight fusion scheme for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 458–471, 2021.

[53] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7254–7263.

[54] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9060–9069.

[55] S. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" 2020, *arXiv:2006.00269*.

[56] Z. Zhang *et al.*, "Bilateral attention network for RGB-D salient object detection," 2020, *arXiv:2004.14582*.

[57] 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 Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2020.2996406.

[58] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[59] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[60] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 451–466.

[61] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1831–1840.

[62] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," 2017, *arXiv:1705.03122*.

[63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[64] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 714–722.

[65] N. Liu, J. Han, and M.-H. Yang, "PicaNet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3089–3098.

[66] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[67] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3085–3094.

[68] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.-Workshops*, 2010, pp. 49–56.

[69] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 280–287.

[70] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey." 2019, *arXiv:1904.09146*.

[71] D.-P. Fan *et al.*, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 186–202.

[72] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum, "One shot learning of simple visual concepts," *Cogn. Sci.*, vol. 33, no. 33, 2011, pp. 1069–7977.

[73] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2017, pp. 4077–4087.

[74] Y. Zeng *et al.*, "Multi-source weak supervision for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6074–6083.

[75] C. Wang, Z. Zha, D. Liu, and H. Xie, "Robust deep co-saliency detection with group semantic," in *Proc. Assoc. Adv. Artif. Intell.*, 2019, pp. 8917–8924.

[76] K. Hsu, Y. Lin, and Y. Chuang, "Weakly supervised saliency detection with a category-driven map generator," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 67.1–67.13.

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

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

[79] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[80] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proc. Int. Conf. Mach. Learn. Workshop*, vol. 2005, Aug. 2005, pp. 74–79.

[81] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[82] Y. N. Dauphin and D. Grangier, "Predicting distributions with linearizing belief networks," 2015, *arXiv:1511.05622*.

[83] M. Feng, H. Lu, and Y. Yu, "Residual learning for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4696–4708, 2020.

[84] L. Zhang *et al.*, "A multistage refinement network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3534–3545, 2020.

[85] X. Qin *et al.*, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.

[86] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6074–6083.

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

[88] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 212–221.

[89] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May. 2004.

[90] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[91] 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.

[92] M. Ran, Z.-M. Lihi, and T. Ayellet, "How to evaluate foreground maps?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.

[93] 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 Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4548–4557.

[94] 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.