HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness

Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, Cédric Demonceaux

HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness

Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma*, and Cédric Demonceaux

***Abstract*—RGB-D saliency detection aims to fuse multi-modal cues to accurately localize salient regions. Existing works often adopt attention modules for feature modeling, with few methods explicitly leveraging fine-grained details to merge with semantic cues. Thus, despite the auxiliary depth information, it is still challenging for existing models to distinguish objects with similar appearances but at distinct camera distances. In this paper, from a new perspective, we propose a novel Hierarchical Depth Awareness network (HiDAnet) for RGB-D saliency detection. Our motivation comes from the observation that the multi-granularity properties of geometric priors correlate well with the neural network hierarchies. To realize multi-modal and multi-level fusion, we first use a granularity-based attention scheme to strengthen the discriminatory power of RGB and depth features separately. Then we introduce a unified cross dual-attention module for multi-modal and multi-level fusion in a coarse-to-fine manner. The encoded multi-modal features are gradually aggregated into a shared decoder. Further, we exploit a multi-scale loss to take full advantage of the hierarchical information. Extensive experiments on challenging benchmark datasets demonstrate that our HiDAnet performs favorably over the state-of-the-art methods by large margins. The source code can be found in https://github.com/Zongwei97/HIDANet/.**

***Index Terms*—Depth-Aware Channel Attention, RGB-D Saliency Detection**

I. Introduction

Salient object detection (SOD) aims to find the most prominent region inside an image that visually attracts human attention. Conventional SOD approaches only take color images as inputs. With deep learning models, RGB SOD has achieved significant success [1]–[5]. However, these models may result in unsatisfactory performance when dealing with complex scenes, e.g., low-contrast light or object occlusion.

Recent advanced RGB-D sensors provide accessibility to depth maps at a low cost. The complementary geometric cues can contribute to scene understanding. In the literature, two main designs have been widely exploited, i.e., single-streaming schemes that combine RGB-D images from the input side [6]–[8] and multi-streaming network that extracts multi-modal features separately and combines them at semantic levels [9]–[16]. Existing networks often directly extract semantic features through the deep network, with few methods fully explore the rich geometric priors provided by the depth map.

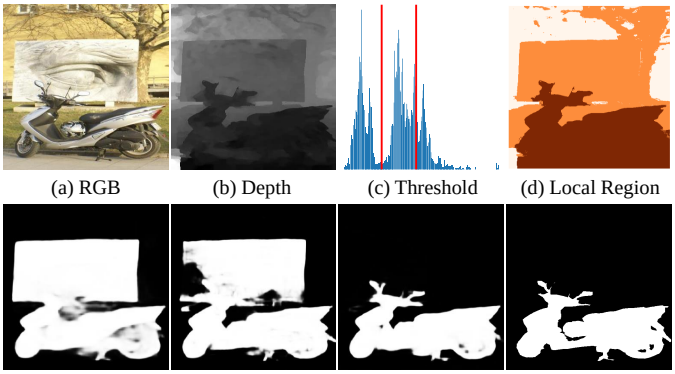
Z. Wu and C. Demonceaux are with ImViA, Université de Bourgogne, Dijon, France.

G. Allibert is with Université Côte d'Azur, CNRS, I3S, Nice, France.

F. Meriaudeau is with ICMUB UMR CNRS 6302, Université de Bourgogne, Dijon, France

Z. Wu and C. Ma are with MOE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China

* Corresponding author is C. Ma: chaoma@sjtu.edu.cn



(a) RGB (b) Depth (c) Threshold (d) Local Region

(e) DASNet (f) SPNet (g) Ours (h) GT

Fig. 1. **Motivation** of our hierarchical depth awareness. (a) and (b) are the paired RGB-D inputs. From the depth histogram, we generate the Multi-Otsu thresholding (c) and divide the input image into local regions. We observe that these regions, referring to the depth granularity, correlate well with the neural network hierarchies. Therefore, we take full advantage of depth priors to improve the feature discriminatory power. Compared to two state-of-the-art (SOTA) RGB-D models (e) and (f) [13], [17], our method favorably yields results (g) closer to the ground-truth mask (h).

Previous works on channel attention [18]–[21] have shown their effectiveness in emphasizing the attentive features among channels. A number of saliency detection works [9], [10], [12], [17] adopt channel attention to enhance multi-modal features. However, the first step of learning channel attention is to aggregate the spatial information of feature maps to construct a $1 \times 1 \times C$ vector by using global average pooling, where C is the number of channels. As a result, the foreground and background contribute equally to the output, which is not optimal to distinguish salient objects. Considering these issues, an intuitive motivation is to design local channel attention referring to depth priors in order to improve feature representation learning.

As shown in Fig. 1, while dealing with complex scenes, current state-of-the-art (SOTA) RGB-D models [13], [17] fail to extract the salient region due to similar visual appearance between the foreground and background (Fig. 1(f) and (g)). However, we observe that salient regions often share similar depth properties, i.e., a certain granularity of depth prior, that help to distinguish the salient objects from the background (Fig.1(b) and (d)). Inspired by this observation, we develop a local feature enhancement scheme with granularity-based attention (GBA) to improve saliency detection. Specifically, we propose to first generate various local regions according to the granularity via Otsu thresholding [22], [23]. These regions can be considered as distinct local spatial attention. Then for each region, we apply local channel attention to improve the

feature discriminatory power. Fig. 1(c) and (d) illustrates such an example of the Otsu threshold values and granularity-aware masks, respectively. We show that our approach can better reason about salient regions (Fig. 1(g)) that are closer to the ground truth (Fig. 1(h)).

We further introduce a cross dual-attention module (CDA) to learn channel and spatial attention from auxiliary modalities to improve the current streaming. The enhanced features are hierarchically fused for final saliency map generation. Besides, the same cross-interaction scheme is embedded to articulate features between encoders and decoders through a U-Net-like [24] architecture. We attentively mirror the multi-scale encoder features to preserve valuable geometric priors within each decoder. The encoded features are gradually fused to a shared decoder. Finally, we use a multi-scale loss on top of outputs from each decoder to optimize the saliency map. Concretely, our contributions are summarized as follows:

- We propose a novel granularity-based attention scheme that attends to fine-grained details in order to strengthen the feature discriminability of each modality.
- We design a new multi-modal and multi-level fusion scheme with a multi-scale loss to take full advantage of the network hierarchy.
- We extensively validate our HiDAnet on large-scale challenging benchmarks. Our approach performs favorably over SOTA models with large margins.

II. Related Work

There are extensive surveys [25]–[30] of salient object detection in the literature. In this section, we briefly review related RGB-D saliency detection as follows:

Multi-Modal Fusion. The auxiliary depth map provides extra geometric clues in addition to visual appearance. To efficiently merge both modalities, several fusion methods have been proposed. A number of works [6]–[8], [31]–[33] directly concatenate the depth map with RGB images from the input side through a single-stream network. On the one hand, JLDCF and its successor [7], [31] explore the siamese design for saliency detection by concatenating RGB and depth images in an additional dimension with a joint learning scheme. DANet [6] forms a four-channel input and enhances the extracted features with a dual-attention mechanism learned from depth. [8], [32] propose the stochastic framework to analyze the uncertainty during human labeling and model the distribution of the saliency output. Different from previous works, [33], [34] attempt to address RGB-D SOD from the 3D point of view with a 3D convolutional neural network. The recent [35] leverages the depth cues to mimicks multi-view images and then fuse them to form the final output.

On the other side, multi-stream models [9]–[16], [36] have achieved leading performances in RGB-D SOD. These models adopt two parallel encoders on different modalities, and the features are fused through different strategies. Several works [9], [37], [38] firstly enhance the depth features before fusing with RGB features. It is worth noting that a portion of the depth maps in existing saliency datasets are not of satisfactory quality. As discussed in [7], [39]–[41], the depth may contain measurement or estimation bias. Thus, DCF [10] designs a calibration module to improve the depth quality. [36], [40]–[42] propose a layer-wise attention to model the geometric contribution with respect to the network depth. [41] explores an additional backbone to learn the weighting scalar purely from depth. [40] analyzes the similarity between RGB and depth features to regular the depth contribution. Sharing the same motivation, [36] computes the reliability of each modality at each stage and then merges them through their reliability. Instead of learning the weighting scalar, [42] generates the weighting maps at each scale to calibrate the feature response. Similarly, [43] leverages bilateral attention to improve foreground-background features separately. Unlike these works, we first divide the feature map into several local regions with the help of depth granularity. The feature maps are further calibrated with different local attention to improve the feature discriminability. Compared to [42], [43], our fined-grained details are statically computed by maximizing the inter-class distance without learning parameters, leading to more reasonable and stable locally-calibrated areas.

There exist other works which only extract features from RGB input while the depth map only serves as supervision [17], [44], [45]. In this context, [46], [47] propose to leverage the pseudo-depth to guide the RGB learning. A2dele [44] further formulates depth supervision as a knowledge transfer problem. CoNet [45] and DASnet [17] propose a multi-task learning framework with an additional depth head together with the saliency branch. However, we argue that these methods cannot fully leverage the multi-modal cues during feature extraction. Instead, we propose a cross-interaction scheme to take full advantage of cross-modal cues. We benefit from the auxiliary modality to alleviate errors in the feature modeling (depth to RGB, and RGB to depth).

Multi-Level Fusion. U-Net with skip connections [24] has shown its effectiveness in pixel-level segmentation tasks. Several RGB-D SOD models [11], [13], [14], [48] equip this design for clearer boundary generation. [48] adopts the feature-wise addition. [13], [14] concatenate the encoder features with the decoder. [11] designs a dense connection between high-level features and the decoder. In this work, we exploit the contribution of attention modules for skip connections applied to SOD. It is worth mentioning the success of skip connections can be mainly attributed to aggregation between the semantic features provided by the contracting path and fine-grained features from the expansion path. From a new perspective, we consider the encoder-decoder features as multi-modal features, and a unified cross-fusion scheme is applied to boost the performance.

Attention for Feature Enhancement. Attention methods such as transformer [49], CBAM [18], SEnet [19], DA [50], and ECA [21] have demonstrated their success in other vision tasks. A number of RGB-D saliency models also equip attention modules to extract attentive features from different modalities. VST [51] and TriTrans [14] adopt transformer [49] for saliency detection. [17], [52], [53] apply the SE module to compute modality-specific attention for feature calibration. Similarly, CDInet [11] designs a depth-induced channel attention to enhance RGB features. From another perspective, [54]

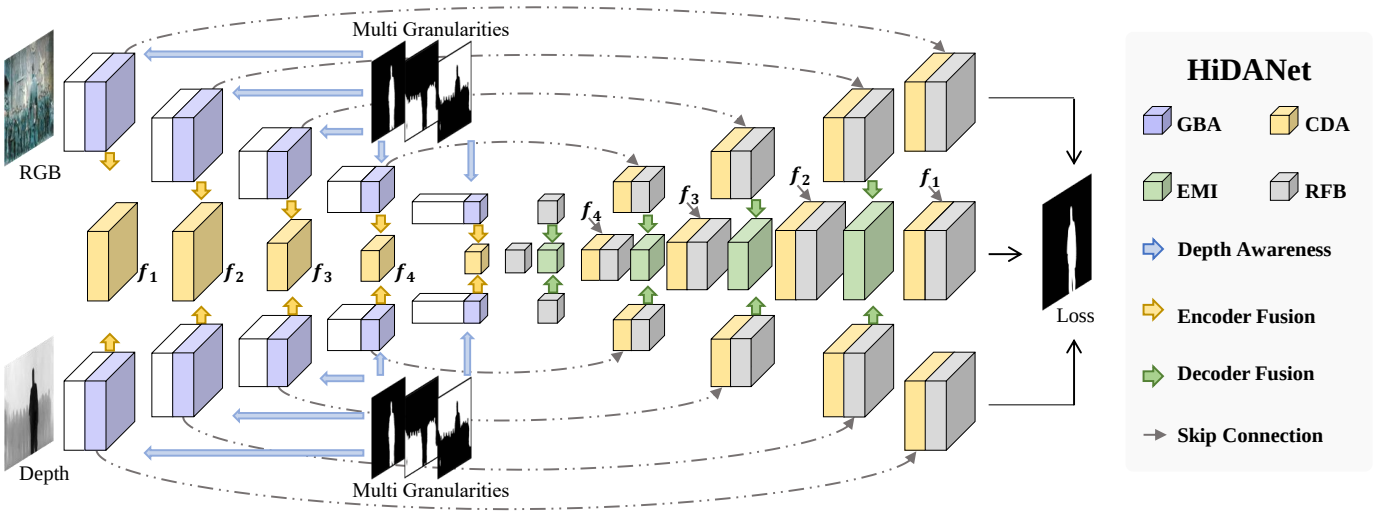


Fig. 2. The overall architecture of our HiDAnet with U-Net-like design. It consists of granularity-based attention (GBA Section III-A), cross dual-attention module (CDA Section III-B), and efficient multi-input fusion (EMI Section III-C). RFB is the receptive field block from [57] for accurate object detection. White blocks denote the network backbone. Our granularity-based attention strengthens the discriminatory power of RGB and depth features separately. Our cross dual-attention module takes advantage of cross-domain cues to attentively realize multi-modal and multi-level fusion in a coarse-to-fine manner. Our efficient fusion scheme effectively models the shared information from each modality. The shared features are further improved with the skip connections for final saliency map generation. Best viewed in color.

deeply explores the spatial attention at different scales with the help of decoupled dynamic convolution. Sharing the same motivation, DFMnet [16] adopts a depth holistic attention on top of features with different resolutions. More recently, several works leverages both spatial and channel attention to jointly improve the feature representation. For example, BBSnet [9] applies the CBAM [18] on the depth map to improve the depth quality before fusion. [55] further improves the CBAM by highlighting spatial features. Sharing the same motivation, CMINet [12] applies the DA [50] on to lately merge RGB-D features. Different from previous works with bi-directional cross-modal attention, HAINet [56] explores the purified depth to improve the RGB features in turn.

Despite the proven effectiveness, previous channel attention schemes do not fully benefit from the geometric priors. For example, the same attention can be applied to both foreground and background. The rich geometric priors in the input depth map have rarely been discovered, which limits the performance of RGB-D saliency detection. DSA2F [15] introduces a depth-sensitive module with the help of the depth histogram. However, it computes the depth region with a fixed threshold for each input image and the attention scores are simply computed by a $Conv_{1\times1}$. In contrast, we propose to dynamically generate multi-granularity regions with the multi-Otsu method [22], [23]. The fine-grained details are further integrated with channel attention to enhance the feature discriminability for sharper edge generation.

III. Method

Fig. 2 presents the overall framework of our proposed HiDAnet. Note that the Otsu masks are generated from the depth map during the pre-processing. Firstly, RGB and depth maps are fed into two parallel encoders for feature extraction. For each individual encoder (RGB/Depth), we propose a granularity-based module (GBA) with the help of input Otsu masks to enhance the discriminatory power, e.g., foreground and background. This module is naturally embedded into different levels of the encoder to correlate with the network hierarchies. With the enhanced features, we propose a unified fusion mechanism (CDA) for multi-modal and multi-level fusion. It enables a cross-domain interaction with both channel and spatial attention to learn the informative shared features in a coarse-to-fine manner. These features are later gradually aggregated into the shared decoder through the efficient multi-input fusion module (EMI). Lastly, we exploit a multi-level loss to take full advantage of the network hierarchies. Details of each component are presented in the following sections.

A. Feature Extraction with Granularity-Based Attention

We observe that the multi-granularity properties of geometric priors correlate well with the network hierarchies of saliency models. Inspired by this observation, we propose the granularity-based attention that aims to attentively combine the spatial attention mask with the conventional channel attention as shown in Fig. 3. For earlier layers, it strengthens the low-level representations to precisely localize the salient object with a sharp boundary. For deeper layers, it improves the semantic abstraction and contributes to the identification of salient objects regardless of appearance variations.

Given the depth map D with its histogram H, we dynamically generate the fine-grained details. According to the value/distance within the depth map, we use the Otsu algorithm [22] to discretize the histogram H into several different regions. In this work, we use the extended multi-Otsu [23] to generate multiple thresholds. Assuming T random thresholds $(d_1, d_2, ..., d_T)$ dividing the depth into $T+1$ parts. Let (σ_i^2, w_i) be the variance and the pixels number of region i $(1 \le i \le T+1)$. The optimal values $\{d_1^*, d_2^*, ..., d_T^*\}$ are chosen by maximizing the inter-class variance:

$$\{d_1^*, d_2^*, ..., d_T^*\} = argmax\{\sigma_w^2(d_1, d_2, ..., d_T)\}, \quad (1)$$

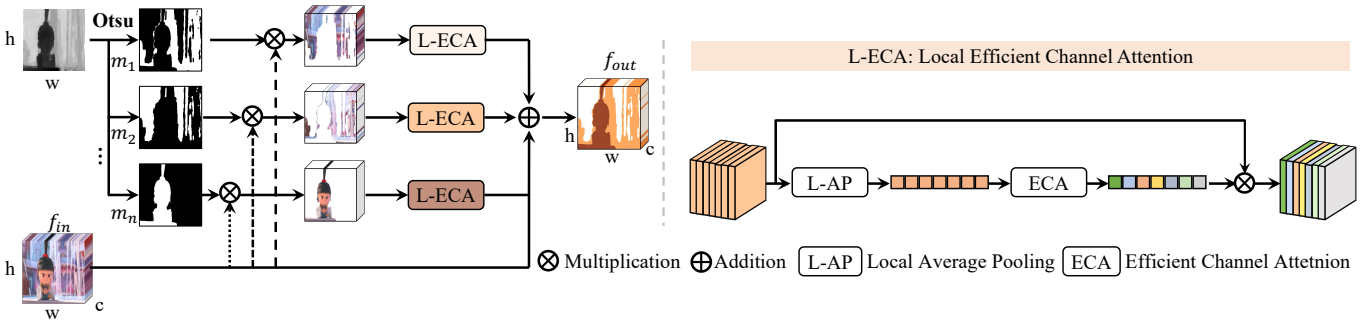


Fig. 3. Diagram of the **granularity-based attention**. The depth awareness is encoded via Local Efficient Channel Attention (L-ECA). ECA is from [21].

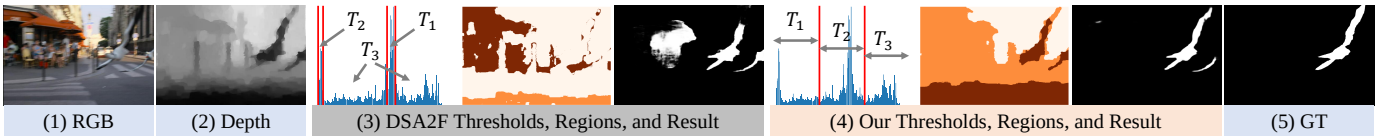


Fig. 4. **Visual comparison** with concurrent DSA2F [15]. DSA2F adopts a recursive thresholding strategy based on the largest depth distribution with pre-defined interval windows. We observe that such thresholding may reason out misleading segmentation without handcraft adjusting. Differently, our method maximizes the inter-class variance, leading to more accurate masks compared to DSA2F across different scenarios. We further explore the granularity cues via channel attention, yielding results closer to the ground truth (5).

where $\sigma_w^2 = \sum_{i=1}^{T+1} w_i \sigma_i^2$. To reduce the computational cost, we only generate the Otsu regions once during pre-processing and further resize them to fit the resolution of feature maps from different scales.

For the i^{th} region m_i, $(1 \leq i \leq T+1, i \in \mathbb{N}^*)$, we mask out the feature map f_{in} with element-wise multiplication to suppress the inactive area through $f_{in} \otimes m_i$. Then, the channel attention is applied to improve the feature representation with local awareness. Compared to the vanilla channel attention [19], [21], we replace the global average pooling with the local average pooling that attends to the local details referring to geometric priors. Finally, the locally enhanced features are aggregated by a residual connection for the final output generation f_{out}. The overall process can be formulated as:

$$\begin{aligned} L-ECA(x) &= \sigma(Conv_{1d}(L-AP(x))) \otimes x, \\ f_{out} &= \sum_{i=1}^{T+1} L-ECA(f_{in} \otimes m_i) + f_{in}, \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is the Sigmoid activation, $\otimes$ is the element-wise multiplication, and $L-AP$ denotes the local average pooling on each masked region. We provide more details on the differences between the proposed granularity-based attention and traditional channel attention in the ablation study Section V Tab. V.

Remarks. Several previous works have proposed to explore depth prior in various manners such as the contrast in CPFP [58], the edge in CoNet [45], or the histogram in DSA2F [15]. Our approach resembles the DSA2F that both methods belong to threshold-based segmentation frameworks. However, one main difference is that we dynamically generate optimized masks by maximizing the inter-class variance, while DSA2F thresholds the histogram with a fixed window size. Specifically, DSA2F targets the largest depth distribution based on which a pre-defined interval window is chosen as the threshold. The process is repeated in a recursive manner to segment the image. However, as shown in Fig. 4, we observe that such a method may fail to accurately measure the granularity scenarios for all the scenes. Differently, our method is based on Otsu thresholding where we can dynamically optimize the thresholds by maximizing inter-class variance without handcraft adjusting. We can observe that our approach computes more discriminative regions, yielding a more effective and robust manner to explore the depth prior. Moreover, thanks to the Otsu optimization, our generated masks are more robust to the depth noise compared to the concurrent work, leading to better preserved local structure with sharpened contour. Finally, we attentively merge the fine-grained details with the semantic cues along the channel attention, while DSA2F simply uses a $Conv_{1\times1}$ for local feature modeling. As shown in Fig. 4, our method can reason about more accurate saliency regions closer to the ground truth. The quantitative comparison with [15], [45], [58] can be found in Section IV-C Tab. I. Our superior performance proves that we can better model the depth priors.

B. Encoder Fusion with Cross Dual-Attention Module

Previous studies [17], [39], [44] have affirmed the effectiveness of learning from two heterogeneous modalities for RGB-D SOD. Color images provide rich information in visual appearance while depth maps contain more spatial priors. Both modalities contribute to modulating homogeneous semantic information. Therefore, the objective of multi-modal learning is to efficiently fuse features with diverse information from different modalities. Similar to multi-modal features, multi-level features also contain both heterogeneous and homogeneous information: high-level features are richer in abstract semantic cues while low-level features are richer in fine-grained details. Thus, from a new perspective, we design a unified fusion scheme to make full use of cross-domain cues for both multi-modal and multi-level reasoning.

Assuming two paired multi-modal features f_x and f_y. We firstly build a transformation F_t to map the inputs $f_x, f_y \in$

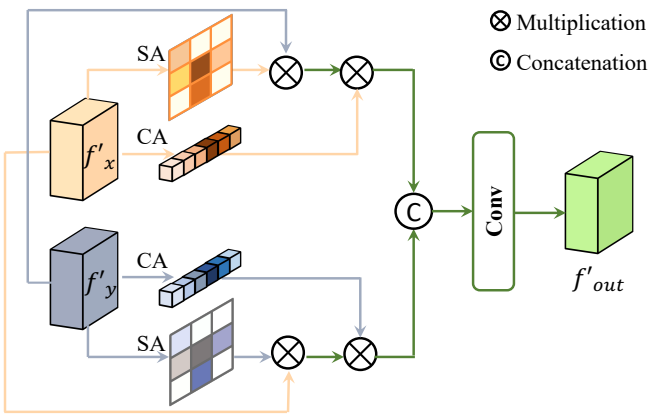


Fig. 5. The proposed multi-scale multi-level **encoder fusion** scheme with cross-domain supervision. Best viewed in color.

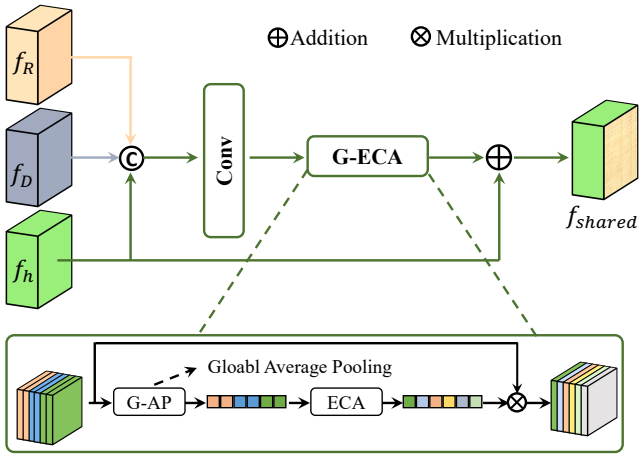


Fig. 6. The proposed efficient **decoder fusion** scheme for multi-type inputs. By fully exploiting the channel-wise dependencies, input features are attentively aggregated to generate the shared output. Best viewed in color.

$\mathbb{R}^{C\times h\times w}$ to feature maps $f'_x, f'_y \in \mathbb{R}^{C'\times h\times w}$ with $C' = \frac{C}{2}$. Specifically, F_t is the combination of a 1×1 convolution which halves the channel size and a 3×3 convolution which is expected to activate the edge response:

$$\begin{aligned} f'_x &= F_t(f_x) = Conv_{3\times 3}(Conv_{1\times 1}(f_x)), \\ f'_y &= F_t(f_y) = Conv_{3\times 3}(Conv_{1\times 1}(f_y)). \end{aligned} \tag{3}$$

Once obtaining the lightweight representation, the next step is to aggregate features from different domains (RGB-D or encoder-decoder). We observe from Fig. 1 that the fine-grained details, such as relative boundary, facilitate the identification of salient objects. Simultaneously, in case it is difficult to distinguish objects at the same distance on the depth map, e.g., when distinguishing the motorbike from the street, the visual appearance becomes more reliable. Inspired by this observation, we aim to use heterogeneous clues to compensate for the single-domain streaming.

To this end, we propose a cross dual-attention fusion scheme as shown in Fig. 5. Specifically, from each input feature map, we learn the 1-D channel attention $M_c \in \mathbb{R}^{C'\times 1\times 1}$ to determine *what* information to be involved, and the 2-D spatial attention $M_s \in \mathbb{R}^{1\times h\times w}$ to determine *which* part to focus. We formally have the operations:

$$\begin{aligned} M_c(f') &= \sigma(MLP(GAP(f')) + MLP(GMP(f'))), \\ M_s(f') &= \sigma(Conv_{7\times 7}(Concat(CAP(f'), CMP(f')))), \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid activation, MLP is the multi-layer perceptron, GAP and GMP are the global average and max pooling, respectively, and CAP and CMP are the average and max pooling across the channel, respectively. With the learned dual attention from separate feature maps, we enable a cross-domain interaction. In such a way, we can alleviate the ambiguities in the domain-specific features. Finally, the cross-enhanced features are fed into concatenation and convolution to form the shared representation f'_{out}. The overall process can be formulated as:

$$\begin{aligned} f_x^{enh} &= M_s(f'_y) \otimes M_c(f'_y) \otimes f'_x, \\ f_y^{enh} &= M_s(f'_x) \otimes M_c(f'_x) \otimes f'_y, \\ f'_{out} &= Conv_{3\times 3}(Concat(f_x^{enh}, f_y^{enh})), \end{aligned} \tag{5}$$

where $\otimes$ denotes element-wise multiplication. For the shared encoder, starting from the second layer, once the multi-modal features are fused through cross attention, the output is further combined with the previous level output through a $Conv_{3\times 3}$.

Remarks. Our fusion design differs from concurrent works [10], [13], [17], [59] in several aspects: (**A**) We leverage both spatial and channel attention to aggregate multi-modal features, while [10], [17] only focus on channels; (**B**) Different from ASTA [59], our calibration is bi-directional (RGB to depth and depth to RGB), while ASTA is asymmetric which only leverages depth cues to improve RGB features. Hence, it does not tackle depth noise; (**C**) SPNet [13] also adopts the symmetric fusion strategies. Our work differs from SPNet in that we fully explore the attention modules for feature fusion, while SPNet is built upon simple convolutions to combine features; (**D**) The fusion scheme can also be implemented by the CBAM [18]. However, vanilla CBAM is modality-specific and cannot explore its relevance in cross-domain features. The ablation study in Section V-C Tab. IX shows the gain with the cross interaction.

C. Decoder Aggregation with Efficient Multi-Input Fusion Module

To aggregate the learned features from both RGB and depth decoders into the shared decoder, a simple concatenation may not be adaptive enough due to the tripled number of descriptors. Thus, we propose an efficient multi-input fusion strategy. Specifically, as shown in Fig. 6, after the simple concatenation between different inputs (RGB f_R, depth f_D, and previous-level shared features f_h), we adopt the vanilla ECA [21] module (termed G-ECA with global pooling) to explore the inter-dependencies of different features. Thus, the most responded features are adaptively selected to form the shared decoder. A residual addition is adapted to reinforce the contribution of the previous level features. We have the overall process:

$$f_{shared} = G-ECA(Conv_{3\times 3}(Concat(f_R, f_D, f_h))) + f_h. \tag{6}$$

The shared decoded features are then fed into our cross dual-attention scheme to realize the skip-connection between the shared encoder-decoder.

Remarks. Our encoder fusion (CDA) and decoder fusion (EMI) are technically different. We observe that the spatial cues are gradually lost during encoding and become limited for decoders. This motivates us to apply both spatial and channel attention for the encoder fusion, while only using channel attention for the decoder fusion.

D. Optimization

To take full advantage of the hierarchical information, we supervise multi-level outputs for both RGB, depth, and shared/fused branches. For outputs from each level, the predicted map is upsampled to form the same resolution mask as the ground truth. We adopt BCE loss $\mathcal{L}^{BCE}$ for pixel restriction and IoU loss $\mathcal{L}^{IoU}$ for global restriction [17], [60], [61]. Therefore, we have the loss $\mathcal{L}_i$ for the i^{th} level output:

$$\mathcal{L}_i = \mathcal{L}_i^{BCE} + \mathcal{L}_i^{IoU}. \tag{7}$$

In total, we have five-level outputs (after each RFB in Fig. 2). Thus, by combining the loss from each branch (R for RGB, D for depth, and S for shared branches), the overall multi-level loss function $\mathcal{L}_{ml}$ becomes:

$$\mathcal{L}_{ml} = \sum_{i=1}^{5} \lambda_i (\mathcal{L}_i(R) + \mathcal{L}_i(D) + \mathcal{L}_i(S)), \tag{8}$$

where λ_i is the weight of the different-level loss. To correlate with the network hierarchies, we follow [17], [62] and set the weight λ as $\{1, 0.8, 0.6, 0.4, 0.2\}$.

We expect the multi-level loss to measure the difference between the generated mask and ground truth at various layers, and to force the network to learn hierarchical features that capture long- and short-range spatial relationships between pixels. The gain by adopting the multi-level loss can be found in the ablation study Section V Tab. VIII.

IV. Experiments

A. Benchmark Datasets

To verify the effectiveness of our approach, we firstly train with the conventional training dataset following the protocol presented in [9], [10], [13], [14], [17] with 2,195 samples: 1,485 samples from the NJU2K-train [63] and 700 samples from the NLPR-train [64]. For testing, experiments are conducted on five classical benchmark RGB-D datasets. DES [65] : includes 135 images of indoor scenes captured by a Kinect camera. NLPR-test [64]: contains 300 natural images captured by a Kinect under different illumination conditions. NJU2K-test [63]: contains 500 stereo image pairs from different sources such as the Internet, 3D movies, and photographs taken by a Fuji W3 stereo camera, where several depth maps are estimated through an optical flow method [66]. STERE [67]: includes 1,000 stereoscopic images downloaded from the Internet where the depth map is estimated using the SIFT flow method [68]. SIP [39]: contains 929 images with humans in the scene, and images are acquired by a mobile device. We further evaluate our model on a newly published dataset COME15K [12] where the depth is estimated through a modified optical flow algorithm [69]. In this case, our model is trained with provided 8,025 training samples and tested on the "Difficult" set with 3,000 images.

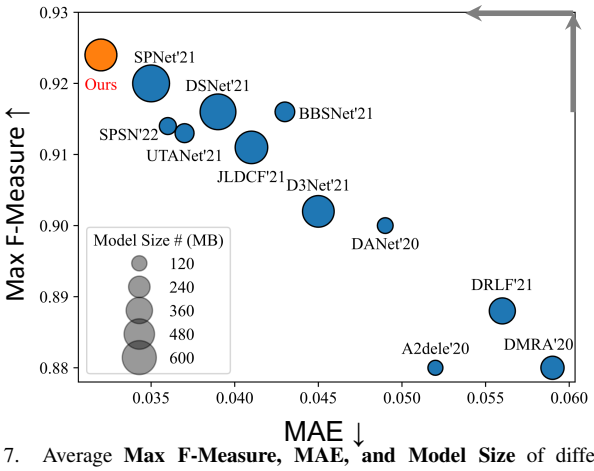


Fig. 7. Average **Max F-Measure, MAE, and Model Size** of different methods on benchmark datasets. The circle size denotes the model size. Note that better models are shown in the upper left corner (i.e., with a larger F-measure and smaller MAE). Methods with smaller size perform inferior, making our method both efficient and accurate.

B. Experimental Settings

Our model is implemented based on Pytorch and trained with a V100 GPU. Our backbone is initialized with the pre-trained weights obtained from ImageNet. For the depth stream, we modify the first convolution to start from one channel. The input RGB-D resolution is fixed to 352×352. We choose the Adam algorithm as our optimizer. We initialize the learning rate to be $1e{-}4$ which is further divided by 10 every 60 epochs. The total training time takes around 6 hours for 100 epochs. During training, we adopt random flipping, rotating, and border clipping for data augmentation. During inference, the prediction maps from the shared branch are the final outputs (middle branch of Fig. 2).

We evaluate our performance with four generally-recognized metrics: F-measure is a region-based similarity metric that takes into account both Precision (P) and Recall (R). Mathematically, we have : $F_\beta = \frac{(1+\beta^2)\cdot P\cdot R}{\beta^2\cdot P+R}$. The value of β^2 is set to be 0.3 as suggested in [77] to emphasize the precision. In this paper, we report the **maximum F-measure** (F_β) score across the binary maps of different thresholds. **Mean Absolute Error** (M) measures the approximation degree between the saliency map and ground-truth map at the pixel level. **S-measure** (S_m) [78] evaluates the similarities between object-aware (S_o) and region-aware (S_r) structures of the saliency map compared to the ground truth. Mathematically, we have: $S_m = \alpha \cdot S_o + (1-\alpha)\cdot S_r$, where α is set to be 0.5. **E-measure** (E_m) evaluates both image-level statistics and local pixel-matching information. Mathematically, we have: $E_m = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi_{FM}(i,j)$, where $\phi_{FM}(i,j)$ stands for the enhanced-alignment matrix as presented in [79]. To make a fair comparison, we use the same protocol as [13] to evaluate the officially released saliency maps for each SOTA method.

TABLE I
QUANTITATIVE COMPARISON WITH SOTA MODELS. ↑ (↓) DENOTES THAT THE HIGHER (LOWER) IS BETTER. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

Dataset	Size	DES				NLPR				NJU2K				STERE				SIP			
Metric	Mb	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
Performance of RGB-D Models with VGG Backbones																					
$DMRA_{19}$ [70]	278	.030	.907	.900	.934	.031	.888	.899	.940	.051	.896	.886	.920	.047	.895	.886	.930	.086	.852	.806	.847
$A2dele_{20}$ [44]	116	.029	.897	.886	.917	.029	.895	.899	.943	.051	.890	.871	.914	.044	.892	.879	.926	.070	.856	.829	.887
$ATSA_{20}$ [59]	131	.022	.931	.917	.954	.027	.907	.909	.947	.046	.905	.885	.928	.038	.912	.896	**.940**	.063	.884	.849	.895
$CMMS_{20}$ [71]	546	.018	.934	.934	.958	.028	.914	.919	.946	.044	.905	.900	.929	.045	.899	.894	.925	.058	.893	.872	.901
$DANet_{20}$ [6]	128	.029	.916	.904	.932	.047	.904	.897	.926	.045	.910	.899	.927	.048	.895	.892	.919	.054	.900	.888	.912
$CMWNet_{20}$ [72]	327	.022	.939	.934	.959	.029	.913	.917	.941	.046	.913	.903	.925	.043	.911	.905	.930	.062	.889	.867	.901
$HDFNet_{20}$ [48]	308	.021	.932	.926	.962	.023	.926	.923	.957	.039	.922	.908	.939	.042	.910	.900	.933	.048	**.909**	.886	.924
$PGAR_{20}$ [73]	62	.032	.894	.886	.906	.027	.912	.917	.941	.042	.918	.909	.932	.045	.902	.894	.919	.072	.852	.838	.875
SSF_{20} [74]	126	.026	.912	.904	.930	.027	.912	.915	.947	.043	.911	.899	.929	.065	.859	.837	.882	.091	.810	.799	.855
$CASGNN_{20}$ [75]	160	.027	.917	.893	.926	.025	.914	.919	.953	.036	.927	.910	.944	**.038**	.913	.899	**.940**	-	-	-	-
$D3Net_{21}$ [39]	518	.031	.909	.897	.923	.030	.907	.912	.942	.049	.910	.900	.928	.039	.911	.902	**.940**	.063	.886	.866	.897
$CDINet_{21}$ [11]	217	.020	.943	**.937**	.962	.024	.923	.927	.953	**.030**	.928	**.918**	.945	.040	.912	**.913**	.937	.054	.904	.875	.908
$UCNet_{21}$ [32]	120	.018	.936	.934	**.970**	.025	.915	.920	.953	.043	.908	.897	.932	.039	.908	.902	.938	.051	.896	.875	.915
$DRLF_{21}$ [76]	351	.030	.909	.895	.918	.032	.904	.903	.929	.055	.896	.886	.914	.050	.897	.888	.916	.071	.869	.850	.882
$HAINet_{21}$ [56]	228	.018	**.945**	.935	.967	.024	.920	.924	.956	.037	.924	.911	.940	.040	**.917**	.907	.938	.052	.907	.879	.917
$BIANet_{21}$ [43]	189	.020	.939	.931	.955	.025	.921	.925	.954	.039	.928	.915	.939	.043	.910	.903	.932	.052	.904	.883	.916
$DCMF_{22}$ [53]	78	.022	.934	.932	.956	.029	.913	.922	.940	.041	.911	.902	.935	.043	.916	.910	.928	-	-	-	-
Ours (VGG16)	269	**.017**	.944	.929	.968	**.021**	**.927**	**.928**	**.962**	.034	**.930**	**.918**	**.947**	.039	.915	.902	.939	**.045**	**.909**	**.889**	**.927**
Performance of RGB-D Models with ResNet Backbones																					
$JLDCF_{21}$ [31]	548	.020	.934	.931	.961	.022	.925	.925	.955	.041	.912	.902	.936	.040	.913	.903	.934	.049	.903	.880	.918
$RD3D_{21}$ [34]	179	.019	.941	.935	.965	.022	.927	.930	.959	.036	.923	.916	.941	.037	.917	.911	.939	.048	.906	.885	.918
$BIANet_{21}$ [43]	244	.020	.939	.930	.958	.023	.924	.926	.956	.036	.929	.917	.942	.039	.912	.905	.935	.047	.904	.887	.920
$CoNet_{20}$ [45]	162	.024	.920	.914	.944	.027	.903	.911	.943	.046	.902	.896	.926	.037	.909	.905	.941	.058	.887	.860	.911
$DASNet_{20}$ [17]	141	.024	.926	.905	.932	.021	**.929**	.929	.960	.042	.911	.902	.935	.037	.915	.910	.939	.051	.900	.877	.918
$BBSNet_{21}$ [38]	200	.021	.942	.934	.955	.023	.927	.930	.953	.035	.931	.920	.941	.041	.919	.908	.931	.055	.902	.879	.910
DCF_{21} [10]	435	.024	.910	.905	.941	.022	.918	.924	.958	.036	.922	.912	.946	.039	.911	.902	.940	.052	.899	.876	.916
$DSA2F_{21}$ [15]	-	.021	.896	.920	.962	.024	.897	.918	.950	.039	.901	.903	.923	.036	.898	.904	.933	-	-	-	-
$DSNet_{21}$ [55]	661	.021	.939	.928	.956	.024	.925	.926	.951	.034	.929	.921	.946	.036	.922	**.914**	.941	.052	.899	.876	.910
$UTANet_{21}$ [52]	186	.026	.921	.900	.932	**.020**	.928	**.932**	**.964**	.037	.915	.902	.945	.033	.921	.910	.948	.048	.897	.873	.925
$C2DFNet_{22}$ [54]	198	.020	.937	.922	.948	.021	.926	.928	.956	-	-	-	-	.038	.911	.902	.938	.053	.894	.782	.911
$MVSalNet_{22}$ [35]	-	.019	.942	.937	**.973**	.022	.931	.930	.960	.036	.923	.912	.944	.036	.921	.913	.944	-	-	-	-
$SPSN_{22}$ [36]	149	.017	.942	.937	**.973**	.023	.917	.923	.956	.032	.927	.918	.949	.035	.909	.906	.941	.043	.910	.891	**.932**
Ours (ResNet50)	523	**.015**	**.947**	**.939**	**.973**	.022	.927	.925	.957	**.030**	**.937**	**.924**	**.952**	**.033**	**.926**	**.914**	**.948**	**.043**	**.915**	**.893**	.930
Performance of RGB-D Models with Res2Net Backbones																					
$BIANet_{21}$ [43]	244	.017	.948	.942	.972	.022	.926	.928	.957	.034	.932	.923	.945	.038	.916	.908	.935	.046	.908	.889	.922
$SPNet_{21}$ [13]	702	.014	.950	.945	**.980**	**.021**	.925	.927	.959	**.028**	.935	.925	**.954**	.037	.915	.907	.944	**.043**	.916	**.894**	**.930**
Ours (Res2Net50)	525	**.013**	**.952**	**.946**	**.980**	**.021**	**.929**	**.930**	**.961**	.029	**.939**	**.926**	**.954**	**.035**	**.921**	**.911**	**.946**	**.043**	**.919**	.892	.927

TABLE II
QUANTITATIVE COMPARISON ON THE CHALLENGING COME15K *Difficult* TEST SET [12]. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

	JLDCF	*A2dele*	*DMRA*	*CoNet*	*BBSnet*	*SPnet*	*CMINet*	**Ours**
$M\downarrow$	.075	.092	.137	.113	.071	.065	.064	**.062**
$E_m\uparrow$	.870	.838	.775	.813	.876	.888	**.893**	**.893**

C. Comparison with SOTA RGB-D models

Quantitative Comparison: We provide in Figure 7 an overview of the average performance on conventional benchmark datasets, i.e., DES [65], NLPR [64], NJU2K [63], STERE [67], and SIP [39]. The detailed quantitative performances can be found in Tab. I. We also present in Tab. II the quantitative comparison on the newly published challenging COME15K [12] dataset. All saliency maps are directly provided by authors or computed by authorized codes.

Under the consideration of a fair comparison, we conduct experiments with different backbones such as VGG16 [80], ResNet50 [20], and Res2Net50 [81]. It can be seen that our HiDAnet with each backbone achieves comparable and superior performance compared to the SOTA models with the same backbone. Specifically, our HiDANet with VGG16 backbones achieves significantly better performance on NLPR and SIP datasets, while being very competitive on the model size with 269 MB and around 6 FPS. Our HiDAnet with ResNet50 backbones further sets new SOTA records on DES, NLPR, and NJU2K datasets with 523 MB and around 12 FPS. We also follow the SOTA SPNet and replace our backbone with Res2Net50. It can be seen that our method performs favorably compared to SPNet with only 525 MB compared to that of SPNet with 702 MB. Our FPS is around 11. We also exhibit in Fig. 9 the PR curves with several latest published models to further demonstrate the superior performance of our model.

Finally, in addition to the difference in the backbone, we observe that existing works adopt different architectures, i.e., design of decoder, supervision, training settings, etc. Under the consideration of fair comparison and to purely analyze the effectiveness of encoder fusion design, we re-

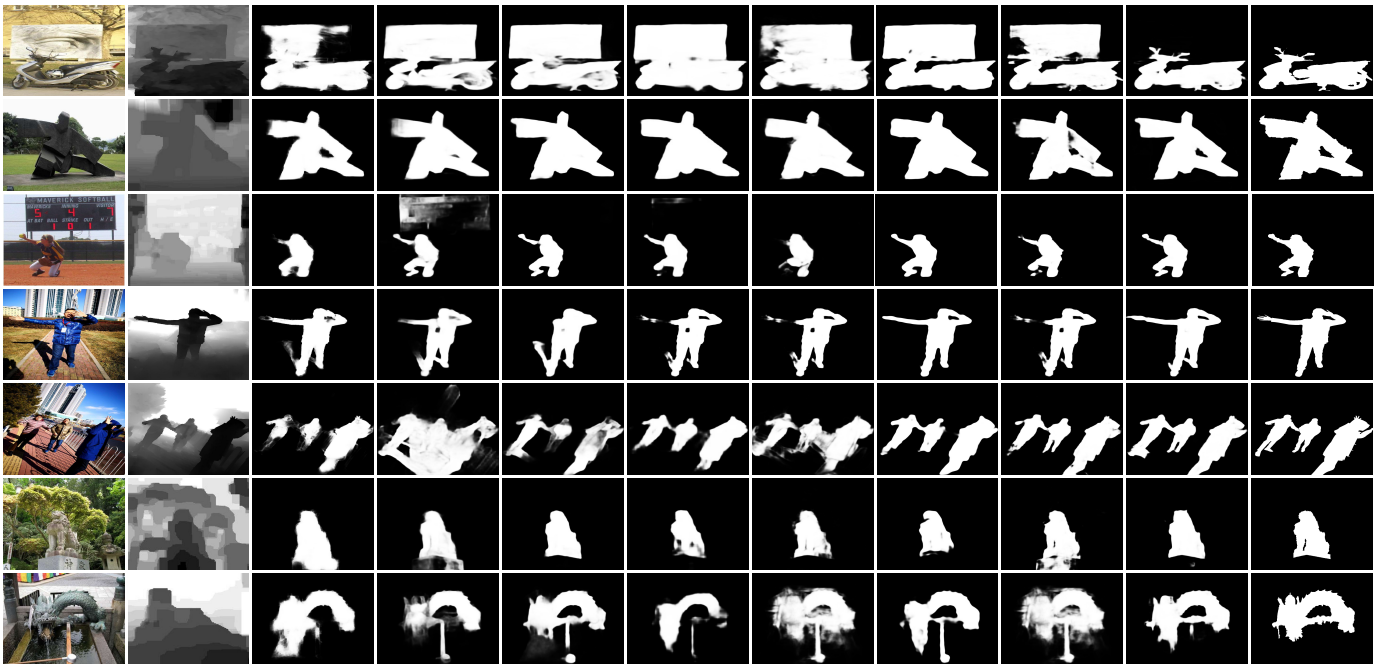


Fig. 8. **Visual comparison** between our HiDAnet and SOTA methods in various challenging cases. It can be seen that our method better explores the granularity prior to reason about the saliency map closer to the ground truth.

TABLE III
QUANTITATIVE COMPARISON WITH DIFFERENT **FUSION DESIGNS**. WE REPLACE OUR FUSION MODULE WITH FOUR SOTA FUSION MODULES AND RETRAIN THE NEW NETWORKS UNDER THE SAME TRAINING SETTING. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

Dataset	Size	NLPR		NJU2K		STERE		SIP	
Metric	Mb	$F_\beta\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$E_m\uparrow$
Res2Net50 + Ours	525	**.929**	**.961**	**.939**	**.954**	.921	**.946**	**.919**	.927
Res2Net50 + BBS [9]	509	.922	.953	.918	.939	.890	.909	.916	.917
Res2Net50 + CDI [11]	531	.926	.958	.927	.946	**.922**	.945	.907	.920
Res2Net50 + DCF [10]	347	.927	.958	.933	.948	.916	.939	.911	.923
Res2Net50 + SP [13]	737	.925	.959	.935	**.954**	.915	.944	.916	**.930**

implement several fusion alternatives under the same architecture (Res2Net50 + fusion). Specifically, we choose the same backbone (Res2Net50), the same decoder (the SOTA [13]), loss (multi-scale supervision), and the same training settings as ours. The only difference between one model to another is in the fusion module. The quantitative comparison can be found in Table III. It can be seen that by replacing our fusion with other methods, the empirical results significantly drop. This validates the superior effectiveness of our granularity and CDA in leveraging RGB-D cues compared to other alternatives.

Qualitative Comparison: Fig. 8 illustrates generated saliency maps of different methods on challenging cases: cluttered background and foreground with a similar appearance ($1^{st}-2^{nd}$ rows), human in the scene ($3^{rd}-5^{th}$ rows), and low contrast on the depth map ($6^{th}-7^{th}$ rows). Compared to the SOTA models, our HiDAnet yields results closer to the ground-truth masks. For the motorbike in the 1^{st} row, our model can selectively remove the background region (board). For the sculpture in the 2^{nd} row, our network pays local attention to the foreground and thus the hollow part can be detailed. We can also accurately extract the human with large deformations ($3^{rd}-5^{th}$ rows).

Robustness against Depth Noise: Tab. IV reports the robustness analysis on the depth quality. To make a fair comparison, we conduct experiments and compare with the SOTA SPnet [13] and CMINet [12] under the same inferior condition with a simulated Gaussian noise on depth. We further evaluate the performances on the simulated noisy testing dataset. The noise level is defined by the conventional metrics *RMSE* and $\delta 1$. While *RMSE* and $\delta 1$ are 0, we report the performance tested with the vanilla dataset (without noise). **Drop** Δ denotes the performance degradation by % under the simulated depth noise.

Note that CMINet designs a multi-scale mutual information minimization during the encoding stage and lately merge multi-modal features at the semantic level, yielding an unsatisfactory performance while dealing with noisy datasets (drop 2.0% S_m and 2.3% E_m for noisy DES). Differently, both SPnet and ours fuse features at each stage, leading to superior robustness against the noise. Compared to SPnet, it can be seen that our performance is more stable, which can be attributed to our granularity attention and fusion designs. The gain of each component can be found in Tab. VIII.

V. ABLATION STUDY

A. Comparison with Vanilla Channel Attention

We propose granularity-based attention (GBA) referring to geometric priors, which differs from the traditional channel attention on the pooling strategies. Formally, let $z \in \mathbb{R}^C$ be the

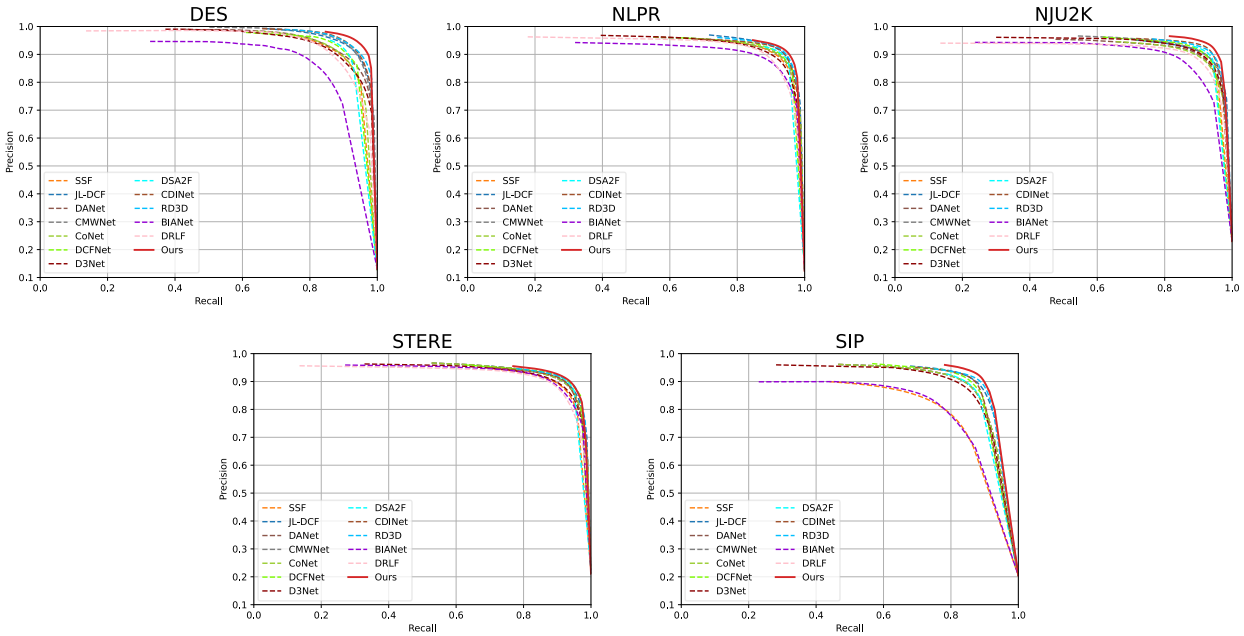


Fig. 9. Comparison on **PR curves**. Our HiDANet achieves better performance compared to the 12 listed SOTA methods across different datasets.

TABLE IV

EXPERIMENTS UNDER **INFERIOR CONDITIONS** WITH SIMULATED DEPTH NOISES (*RMSE*, $\delta 1$). WHILE *RMSE*, $\delta 1$ ARE 0, IT REPRESENTS THE RESULT WITHOUT SIMULATED NOISES. **DROP** Δ DENOTES THE ABSOLUTE PERFORMANCE DIFFERENCE. OUR HIDANET LEADS TO A MORE STABLE PERFORMANCE COMPARED TO THE SOTA METHODS WITH A LOWER Δ UNDER DIFFERENT INFERIOR CONDITIONS, PROVING THAT OUR MODEL IS MORE ROBUST AGAINST DEPTH NOISES. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

Dataset	DES						NLPR						NJU2K					
Metric	*RMSE*	$\delta 1$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	*RMSE*	$\delta 1$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	*RMSE*	$\delta 1$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
CMINet$_{21}$	0	0	.016	.944	.940	.975	0	0	.020	.931	.932	.959	0	0	.028	.940	.929	.954
CMINet$_{21}$	.261	.270	.022	.925	.920	.952	.259	.342	.021	.929	.932	.960	.236	.413	.032	.934	.922	.948
Drop $\Delta(\%)$	-	-	.6	1.9	2.0	2.3	-	-	.1	0.2	**0**	**.1**	-	-	0.4	0.6	.7	.6
SPNet$_{21}$	0	0	.014	.950	.945	.980	0	0	.021	.925	.927	.959	0	0	.028	.935	.925	.954
SPNet$_{21}$	.261	.270	.017	.944	.935	.972	.259	.342	.020	.922	.924	.956	.236	.413	.033	.931	.920	.946
Drop $\Delta(\%)$	-	-	.3	.6	1	.8	-	-	.1	.3	.3	.3	-	-	.5	**.4**	.5	.8
Ours	0	0	.013	.952	.946	.980	0	0	.021	.929	.930	.961	0	0	.029	.939	.926	.954
Ours	.261	.270	.015	.948	.943	.980	.259	.342	.021	.930	.930	.962	.236	.413	.029	.935	.925	.953
Drop $\Delta(\%)$	-	-	**.2**	**.4**	**.3**	**0**	-	-	**0**	**.1**	**0**	**.1**	-	-	**0**	**.4**	**.1**	**.1**

TABLE V

ABLATION STUDY ON ATTENTION DESIGNS WITH **DIFFERENT AVERAGE POOLING** METHODS. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

#	*Description*	DES		NLPR		NJU2K		STERE	
		$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
I	Vanilla Global Attention + Global Pooling	.019	.940	**.020**	**.929**	.030	.936	.037	.918
II	Local Attention + Global Pooling	.015	.947	.021	.927	.032	.928	.038	.915
III	Our Local Attention + Local Pooling	**.013**	**.952**	.021	**.929**	**.029**	**.939**	**.035**	**.921**

squeezed spatial information from feature $x \in \mathbb{R}^{H\times W\times C}$. Accordingly, we can obtain three variations of average pooling:

$$
\begin{aligned}
&(I)\ z = \frac{\sum\sum x(.)}{H\times W};\\
&(II)\ z = \frac{\sum\sum x(.)\cdot m_i()}{H\times W};\\
&(III)\ z = \frac{\sum\sum x(.)\cdot m_i(.)}{\sum\sum m_i(.)}
\end{aligned}
\tag{9}
$$

where (I) denotes the vanilla global average pooling, (II) is the global pooling with local region $m_i(.)$, and (III) is our proposed GBA module that applies local pooling with local region $m_i(.)$.

Note that when depth data is constant, i.e., all the pixels belong to the same granularity, our local average becomes the global average pooling and our model is equivalent to the conventional channel attention [19], [21]. To verify our effectiveness, we conduct experiments by replacing our local pooling with the aforementioned poolings. Empirical results in Tab. V show that compared to (I), (II) can better leverage local awareness which spatially constrains attention around the local region. However, with a large $H\times W$, the attention activation is limited. Hence, we further propose to adopt local pooling to automatically adjust the weight (III). Our superior performance validates the effectiveness of our local design.

TABLE VI
EXPERIMENTS BY GRADUALLY ADDING **GRANULARITY ATTENTION** MODULE ON RGB AND DEPTH STREAMS. RGB(G)/D(G) DENOTES THE CASE WHEN GRANULARITY ATTENTION IS APPLIED TO RGB/DEPTH BRANCH. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

Dataset	DES				NLPR				NJU2K				STERE				SIP			
Metric	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
(A) RGB + D	.015	.949	.940	.972	.022	.925	.927	.960	.030	.932	.923	.952	.037	.913	.901	.936	.046	.914	.889	.923
(B) RGB(G) + D	.014	.951	.943	**.980**	**.021**	.927	.926	.960	.030	.936	.923	.953	.036	.916	.907	.945	**.043**	**.919**	**.894**	**.928**
(C) RGB(G) + D(G)	**.013**	**.952**	**.946**	**.980**	**.021**	**.929**	**.930**	**.961**	**.029**	**.939**	**.926**	**.954**	**.035**	**.921**	**.911**	**.946**	**.043**	**.919**	.892	.927

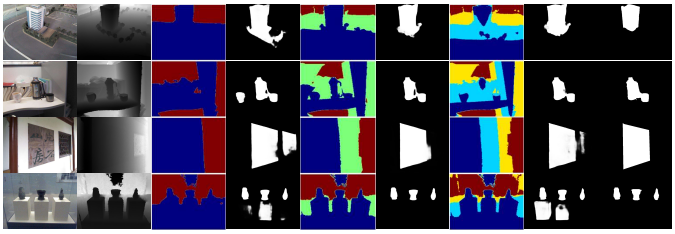


Fig. 10. Qualitative comparison with **different numbers of Otsu** thresholds ($T=1,2,3$) for our granularity-based attention. With the threshold T, we divide the depth map into $T+1$ regions with different colors. Each region shares the same granularity of geometric information. With one threshold $T=1$, the local regions are coarse and cannot get the full benefit from the geometric priors. This results in unsatisfactory salient masks (4^{th} column). With two thresholds $T=2$, the depth map is better discretized with more fine-grained details, yielding salient masks closer to the ground truth (6^{th} column). With three thresholds $T=3$, the depth map is over-discretized, resulting in sub-optimal salient masks (8^{th} column). Our plain HiDAnet is built upon $T=2$.

B. Why GBA in both streams

We analyze in Tab. VI the contribution of GBA for both RGB and Depth feature modelings: (A) We remove the GBA from our network, denoted as RGB+D; (B) GBA is only applied in the RGB stream, denoted as RGB(G) + D; (C) GBA is applied in both streams, denoted as RGB(G) + D(G). We observe that the performance augments by gradually inserting GBA into the encoders. This shows that GBA can be considered as depth-aware attention for the RGB stream and as a self-enhancement module for the Depth stream to produce regions with favorable objectness.

C. Number of Otsu Regions for GBA

Our fine-grained details are determined by the number of Otsu regions as shown in Figure 10. The two first columns represent the paired RGB-D inputs. On the 3^{rd}, 5^{th}, and 7^{th} columns we list the Otsu regions with different numbers of multi granularities, respectively. On the 4^{th}, 6^{th}, and 8^{th} columns we list the generated masks with different numbers of thresholds $T=1,2,3$, respectively.

By comparing the 3^{rd} and 5^{th} columns, it can be seen that a small number of Otsu threshold $T=1$ cannot get the full benefit from the geometric priors. For example, the building in the 1^{st} row cannot be perfectly distinguished from the background; the cups in the 2^{nd} row are mixed with the table and a part of the wall. The unsatisfactory thresholding on the depth histogram leads to sub-optimal performance of granularity-based attention that the discriminatory power cannot be fully exploited. While augmenting the number of thresholds to $T=2$, we observe from the 5^{th} column that the scene can be better discretized. The fine-grained details contribute to the clearer boundary generation as shown in the 6^{th} column. We further augment the number of thresholds to $T=3$ and observe the over-discretization, leading to the misunderstanding on the depth map. Thus, it results in lower quality salient masks as shown in the 8^{th} column.

Thus, we perform the experiments with different numbers of thresholds T. Tab. VII shows that the best overall performance is achieved with $T=2$ thresholds, thus with $n=3$ regions. It can be considered as a scene discretization into three parts: close, middle, and far regions. Our plain HiDAnet is with $T=2$ thresholds and achieves the best performance. We also discover that the sensitivity to thresholding varies from one dataset to another, especially the NLPR dataset which is not highly sensitive to the granularity. This is mainly due to the fact that NLPR contains objects residing in the background. In such circumstances, the target object has the mixed depth response as the background, leading to less-noticeable granularity as shown in the last two rows of Figure 10. In more common and popular cases (DES, NJU2K, STERE, and SIP), our fine-grained details achieve significant improvement compared to our baseline with conventional attention as shown in Tab. VII.

D. Ablation study on Key Components

Tab. VIII presents a thorough ablation study for each key component. We observe that by gradually adding proposed modules, our network leads to better performance. We also conduct experiments by replacing our proposed modules with several SOTA counterparts. Specifically, we compare our Granularity-Based Attention with the DEDA module proposed in [6]. Both our GBA and DEDA belong to the mask-guided attention modules. Specifically, DEDA leverages the depth map to dynamically learn the masked-guided attention map which is supervised by the ground truth. The learned attention map refers to the contrast to guide RGB learning. Differently, our mask is statically computed by the Otsu threshold by maximizing inter-class variance. The computed local regions refer to the fine-grained details which are further integrated with semantics cues. Empirically, by comparing (#6 – #8), our GBA performs favorably against DEDA, showing that our method can better leverage the depth cues to distinguish objects with different camera distances. We also replace our encoder fusion (CDA) with the concurrent DCF [10] built upon channel attention. The main difference is that DCF is based on channel attention, while our CDA additionally leverages the spatial attention for better localization. By comparing

TABLE VII
QUANTITATIVE COMPARISON WITH **DIFFERENT OTSU THRESHOLDS**. OUR PLAIN HIDANET IS WITH $T = 2$ THRESHOLDS. $T = 2$ ACHIEVES THE BEST PERFORMANCE WITH A REASONABLE FPS. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

Dataset		DES				NLPR				NJU2K				STERE				SIP			
Metric	$FPS\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
$T=0$	**13.3**	.019	.941	.927	.955	.020	**.929**	**.931**	**.961**	.031	.936	.924	.952	.037	.919	.908	.943	.046	.915	.888	.924
$T=1$	12.6	.015	.951	**.948**	.979	.023	.927	.927	.960	**.029**	.933	.924	.953	.035	.918	.908	.944	.044	.918	**.894**	**.927**
$T=2$	11.3	**.013**	**.952**	.946	**.980**	.021	**.929**	.930	**.961**	**.029**	**.939**	**.926**	**.954**	**.035**	**.921**	**.911**	**.946**	**.043**	**.919**	.892	**.927**
$T=3$	10.5	.015	.949	.942	.979	**.020**	.929	.928	**.961**	.031	.929	.920	.949	.036	.914	.900	.940	.044	.916	.891	.925

TABLE VIII
ABLATION STUDY ON **KEY COMPONENTS** OF OUR PROPOSED HIDANET. WE PARTIALLY REMOVE KEY COMPONENTS OR REPLACE THE FUSION DESIGNS WITH A SIMPLE ADDITION. *Skip* STANDS FOR THE SKIP CONNECTION WITH THE PROPOSED CROSS DUAL ATTENTION. $\mathcal{L}_{ml}$ DENOTES THE MULTI-LEVEL SUPERVISION. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST; UNDERLINE: SECOND BEST.)

#	Baseline	*GBA*	*CDA*	*Skip*	*EMI*	$\mathcal{L}_{ml}$	*DEDA* [6]	*DCF* [10]	DES		STERE	
									$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
1	✓								.018	.941	.038	.917
2	✓	✓							.016	.944	.037	.917
3	✓	✓	✓						.016	.946	.036	.919
4	✓	✓	✓	✓					.015	.947	.036	**.923**
5	✓	✓	✓	✓	✓				.014	.949	**.034**	.921
6	✓			✓	✓	✓	✓		.016	.946	.041	.914
7	✓	✓		✓	✓	✓		✓	.017	.946	.037	.918
8	✓	✓	✓	✓	✓	✓			**.013**	**.952**	.035	.921

TABLE IX
ABLATION STUDY ON **ENCODER FUSION** AND **DECODER FUSION** DESIGNS. WE USE THE MEAN ABSOLUTE ERROR (M), MAX F-MEASURE (F_m), S-MEASURE (S_m), AND MAX E-MEASURE (E_m) AS EVALUATION METRICS. (**BOLD**: BEST.)

	Configuration	DES		STERE	
		$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
	HiDAnet	**.013**	**.952**	**.035**	**.921**
C1	Add	.017	.945	.039	.915
C2	Cat + Conv	.016	.946	.039	.916
C3	Self + Add	.015	.948	.036	.918
C4	Self + Cat + Conv	.014	.949	.037	.917
C5	Cross + Add	.015	.947	.036	.919
E1	Cat + Conv	.015	.947	.038	.914
E2	Cat + ECA + Conv	.016	.945	.038	.915
E3	Cat + Conv + ECA	.015	.949	.037	.916
E4	E2 + Residual	.014	.950	.036	.920

(#7 − #8), we can observe that while CDA is replaced by the DCF, the performance drops significantly. This validates the effectiveness of our CDA with both channel and spatial attention.

Design of Cross Dual Attention: We verify in Tab. IX the design of our encoder fusion by removing or replacing each component: ($C1$) Features are simply fused through addition; ($C2$) Features are fused through concatenation-convolution (**CC**); ($C3$) Features are firstly self-enhanced with vanilla CBAM before the addition fusion. ($C4$) Features are firstly self-enhanced and later fused through **CC**. ($C5$) We explore the attention in a cross manner and fuse features with addition. We can observe the gain of attention modules by comparing ($C1-C3-C5$), the improvement from cross-domain interaction by comparing ($C3-C5$), and the contribution of **CC** by comparing ($C5-Ours$). These results validate the effectiveness of our proposed encoder fusion scheme.

Design of Efficient Multi-Input Fusion: We also verify the design of our decoder fusion in Tab. IX: ($E1$) Features are fused with **CC**. ($E2$) Features are concatenated and fed into the ECA model before the convolution. ($E3$) Features are fused with CC and then fed into the ECA. ($E4$) Based on the configuration $E2$, we further add a residual addition. By comparing ($E2-E3$) and ($E4-Ours$), we can observe that the ECA module performs better with a reduced channel size. The comparison between ($E2-E4$) validates the effectiveness of residual addition which propagates the hierarchical features.

VI. CONCLUSION

In this paper, we propose an end-to-end HiDAnet for RGB-D saliency detection. Different from previous networks, we fully leverage fine-grained details and merge them with semantic cues through the local channel attention. Extensive evaluations on challenging RGB-D benchmarks indicate that our HiDAnet improves saliency detection in several challenging scenarios where the SOTA approaches fail, notably in cases where multiple objects with similar appearances but at distinct camera distances (granularity). Our method has the potential to be used in many other tasks, including semantic and instant segmentation.

ACKNOWLEDGEMENTS

We gratefully acknowledge Zhuyun Zhou, Jilai Zheng, Renato Martins and Jingjing Wang for discussion and proofreading. Part of the calculations were performed using HPC resources from DNUM-CCUB (Université de Bourgogne). This research is supported in part by the French National Research Agency through ANR CLARA (ANR-18-CE33-0004), the French ADVANCES project ISITE-BFC project (ANR-15-IDEX-0003), and is financed by the French Conseil Régional de Bourgogne-Franche-Comté.

REFERENCES

[1] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[2] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[3] 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 *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge Guidance Network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[6] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[7] 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 *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[10] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated RGB-D salient object detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] C. Zhang, R. Cong, Q. Lin, L. Ma, F. Li, Y. Zhao, and S. Kwong, "Cross-modality discrepant interaction network for RGB-D salient object detection," in *Poceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[12] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao, "RGB-D saliency detection via cascaded mutual information minimization," in *Poceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[13] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving RGB-D saliency detection," in *Poceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[14] Z. Liu, W. Yuan, Z. Tu, Y. Xiao, and B. Tang, "TriTransNet: RGB-D salient object detection with a triplet transformer embedding network," *Poceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[15] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] W. Zhang, G.-P. Ji, Z. Wang, K. Fu, and Q. Zhao, "Depth quality-inspired feature manipulation for efficient RGB-D salient object detection," in *Poceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[17] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" in *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] W. Qilong, W. Banggu, Z. Pengfei, L. Peihua, Z. Wangmeng, and H. Qinghua, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *The IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics (TSMC)*, vol. 9, no. 1, pp. 62–66, 1979.

[23] P.-S. Liao, T.-S. Chen, P.-C. Chung *et al.*, "A fast algorithm for multilevel thresholding," *Journal of Information Science and Engineering (JISE)*, vol. 17, pp. 713–727, 2001.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2015.

[25] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2021.

[26] X. Shen, "A survey of object classification and detection based on 2d/3d data," *arXiv:1905.12683*, 2019.

[27] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, vol. 30, no. 11, pp. 3212–3232, 2019.

[28] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 29, no. 10, pp. 2941–2959, 2018.

[29] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media (CVMJ)*, vol. 5, no. 2, pp. 117–150, 2019.

[30] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: A survey," *Computational Visual Media (CVMJ)*, pp. 1–33, 2021.

[31] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2021.

[32] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[33] Q. Chen, Z. Zhang, Y. Lu, K. Fu, and Q. Zhao, "3-d convolutional neural networks for rgb-d salient object detection and beyond," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2022.

[34] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "Rgb-d salient object detection via 3d convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[35] J. Zhou, L. Wang, H. Lu, K. Huang, X. Shi, and B. Liu, "Mvsalnet: Multi-view augmentation for rgb-d salient object detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[36] M. Lee, C. Park, S. Cho, and S. Lee, "Spsn: Superpixel prototype sampling network for rgb-d salient object detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[37] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2019.

[38] Y. Zhai, D.-P. Fan, J. Yang, A. Borji, L. Shao, J. Han, and L. Wang, "Bifurcated backbone strategy for rgb-d salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 8727–8742, 2021.

[39] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems (TNNLS)*, vol. 32, no. 5, pp. 2075–2089, 2021.

[40] Z. Wu, S. Gobichettipalayam, B. Tamadazte, G. Allibert, D. P. Paudel, and C. Demonceaux, "Robust rgb-d fusion for saliency detection," *2022 International Conference on 3D Vision (3DV)*, 2022.

[41] X. Cheng, X. Zheng, J. Pei, H. Tang, Z. Lyu, and C. Chen, "Depth-induced gap-reducing network for rgb-d salient object detection: An interaction, guidance and refinement approach," *IEEE Transactions on Multimedia (TMM)*, 2022.

[42] N. Huang, Y. Luo, Q. Zhang, and J. Han, "Discriminative unimodal feature selection and fusion for rgb-d salient object detection," *Pattern Recognition (PR)*, vol. 122, p. 108359, 2022.

[43] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for rgb-d salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 1949–1961, 2021.

[44] 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 *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[45] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

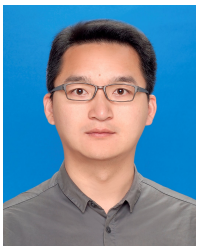
[46] Z. Wu, G. Allibert, C. Stolz, C. Ma, and C. Demonceaux, "Modality-guided subnetwork for salient object detection," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021.

[47] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "Cdnet: Complementary depth network for rgb-d salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 3376–3390, 2021.
[48] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[50] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[51] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
[52] Y. Zhao, J. Zhao, J. Li, and X. Chen, "Rgb-d salient object detection with ubiquitous target awareness," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 7717–7731, 2021.
[53] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for rgb-d saliency detection," *IEEE Transactions on Image Processing (TIP)*, vol. 31, pp. 1285–1297, 2022.
[54] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection," *IEEE Transactions on Multimedia (TMM)*, 2022.
[55] H. Wen, C. Yan, X. Zhou, R. Cong, Y. Sun, B. Zheng, J. Zhang, Y. Bao, and G. Ding, "Dynamic selective network for rgb-d salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 9179–9192, 2021.
[56] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 3528–3542, 2021.
[57] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
[58] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[59] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate rgb-d saliency detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[60] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[61] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
[62] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
[63] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014.
[64] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: a benchmark and algorithms," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.
[65] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS)*, 2014.
[66] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
[67] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
[68] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 5, pp. 978–994, 2011.
[69] J. Wang, Y. Zhong, Y. Dai, K. Zhang, P. Ji, and H. Li, "Displacement-invariant matching cost learning for accurate optical flow estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[70] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
[71] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[72] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020.
[73] S. Chen and Y. Fu, "Progressively guided alternate refinement network for rgb-d salient object detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020.
[74] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020.
[75] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, "Cascade graph neural networks for rgb-d salient object detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020.
[76] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-level recombination and lightweight fusion scheme for rgb-d salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 458–471, 2020.
[77] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[78] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
[79] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.
[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, (ICLR)*, 2015.
[81] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 2, pp. 652–662, 2021.

Zongwei Wu is a PostDoc researcher at the University of Burgundy, France. He received the french engineering degree, equivalent to a master's degree, in mechanical engineering from the University of Technologies of Compiègne - Sorbone university alliance in 2019, and the Ph.D. degree in image processing from the University of Burgundy in 2022. His research interests include computer vision and machine learning. He was a visiting Ph.D. student at Shanghai Jiao Tong University, China, in 2021, and at ETH Zurich, Switzerland, in 2022.

Guillaume Allibert received the M.Sc. degree from the University of Picardie Jules Verne, Amiens, France, in 2005, the Ph.D. degree in automatic control from the University of Orleans, Orléans, France, in 2008 and a "Habilitation Diriger des Recherches" from Université Côte d'Azur, France in 2020. He is currently Associate Professor since 2009 with the Université Côte d'Azur, CNRS, I3S, France. His research interests include computer vision, robotics, nonlinear control and nonlinear observer with applications to unmanned aerial and underwater vehicles.

Fabrice Meriaudeau received the master's degree in physics from Dijon University, Dijon, France, the Engineering degree (with Hons.) in material sciences, in 1994, and the Ph.D. degree in image processing from Dijon University, in 1997. He held a Postdoctoral position with The Oak Ridge National Laboratory, TN, USA in 1997 and several years as a visiting scholar. From 2006 to 2010, he coordinated an Erasmus Mundus Master in the field of computer vision and robotics. From 2010 to 2012, he was the Vice President for International Affairs with the University of Burgundy, Dijon, France. From 2011 to 2016, he was the Director of the Le2i (UMR CNRS), France, which had more than 200 staff members. From 2017 to 2018, he was the Director of the Institute Health and Analytics, Universiti Teknologi PETRONAS Malaysia. He is currently a Professeur des Universites with the University of Burgundy, Dijon, France, ICMUB/IFTIM Research Group in medical imaging. He has authored or co-authored more than 160 international publications and holds three patents. His research interests include image processing for non-conventional imaging systems and more recently medical/biomedical imaging.



Chao Ma is an associate professor at Shanghai Jiao Tong University, Shanghai, 200240, P. R. China. He received the Ph.D. degree from Shanghai Jiao Tong University in 2016. His research interests include computer vision and machine learning. He was a research associate with the School of Computer Science at The University of Adelaide from 2016 to 2018. He was sponsored by China Scholarship Council as a visiting Ph.D. student at the University of California at Merced from the fall of 2013 to the fall of 2015. He is a member of the IEEE.

Cédric Demonceaux received the M.S. degree in Mathematics in 2001 and the PhD degree in Image Processing from the Universite de Picardie Jules Verne (UPJV), France, in 2004. In 2005, he became an associate professor at MIS-UPJV. From 2010 to 2014, he has been a CNRS-Higher Education chair at Le2I UMR CNRS, Universite de Bourgogne. Since 2014, he is a full Professor at the University of Burgundy. His research interests are in image processing, computer vision, and robotics.