Modality-Guided Subnetwork for Salient Object Detection
Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, Cédric Demonceaux

To cite this version:
Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, Cédric Demonceaux. Modality-Guided Subnetwork for Salient Object Detection. 9th International Conference on 3D Vision, Dec 2021, Online, France. hal-03375042v2

HAL Id: hal-03375042
https://hal.science/hal-03375042v2
Submitted on 25 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Modality-Guided Subnetwork for Salient Object Detection

Zongwei Wu¹,³ Guillaume Allibert ² Christophe Stolz¹ Chao Ma³ Cédric Demonceaux ⁴ *
¹ VIBOT EMR CNRS 6000, ImViA, Université Bourgogne Franche-Comté, France
² Université Côte d’Azur, CNRS, I3S, France
³ MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
⁴ ImViA, Université Bourgogne Franche-Comté, France
zongwei_wu@etu.u-bourgogne.fr, allibert@i3s.unice.fr, chaoma@sjtu.edu.cn
{christophe.stolz,cedric.demonceaux}@u-bourgogne.fr

Abstract

Recent RGBD-based models for saliency detection have attracted research attention. The depth clues such as boundary clues, surface normal, shape attribute, etc., contribute to the identification of salient objects with complicated scenarios. However, most RGBD networks require multi-modalities from the input side and feed them separately through a two-stream design, which inevitably results in extra costs on depth sensors and computation. To tackle these inconveniences, we present in this paper a novel fusion design named modality-guided subnetwork (MGSnet). It has the following superior designs: 1) Our model works for both RGB and RGBD data, and dynamically estimates depth if not available. Taking the inner workings of depth-prediction networks into account, we propose to estimate the pseudo-geometry maps from RGB input — essentially mimicking the multi-modality input. 2) Our MGSnet for RGB SOD results in real-time inference but achieves state-of-the-art performance compared to other RGB models. 3) The flexible and lightweight design of MGS facilitates the integration into RGBD two-streaming models. The introduced fusion design enables a cross-modality interaction to enable further progress but with a minimal cost.

1. Introduction

In the last decade, RGB-based deep learning models for salient object detection (SOD) [33, 5, 19, 32, 35] achieved significant success thanks to the advances of GPU and CNN. Given an input image, the goal of SOD is to compute the pixel-wise location of the prominent objects that visually attract human attention the most. However, RGB SOD models focus more on photometric information instead of geometry. This is due to the fixed shape and size kernel design of CNN that is not invariant to scale changes and to 3D rotations. By the lack of geometric information on the input side, it is inevitable for RGB models to add additional learning modules in the network to attend to salient objects, resulting in model complexity and computational cost.

Recent RGBD-based SOD has motivated research interest thanks to the accessibility of cross-modal information from the input side. State-of-the-art RGBD models [9, 26, 23, 36] achieve superior performance over the RGB baseline, affirming the effectiveness of learning from two modalities. Most architectures adapt fusion-wise models, such as early fusion [36] where the depth map is fed as
the fourth channel to RGB image, or multi-scale and late fusion [23] where two-stream networks are adopted. However, early fusion contains more low-level features than semantic ones. Multi-scale or late fusion inevitably requires more learning parameters. As shown in Figure 1, the size of RGBD models is often larger than that of RGB networks.

We explore differently the relationship between depth map and RGB image. Taking human beings as an example, to distinguish salient objects from the 3D world, the input is the visual appearance through human eyes. With the color information and thanks to the depth estimation capability, humans further discover geometric information. This prior guides the understanding of RGB images. It should be the same case for intelligent machines.

To this end, we propose a novel Modality-Guided Subnetwork (MGSnet) which adaptively transforms convolutions by fusing information from one modality to another (e.g., depth to RGB or RGB to depth). Our network matches perfectly both RGB and RGB-D data and dynamically estimates depth if not available by simply applying an off-the-shelf depth prediction model. We design a subnetwork mechanism alongside the master streaming pipeline. The subnetwork can be treated like a light residual-addition branch as the ResNet [10]. It takes one modality map as the master input, e.g. RGB, and enhances its robustness by deforming the convolution kernel with the supervision of the complementary modal prior, e.g. depth, and vice versa.

In summary, the main contributions of this paper are listed as follows:

- By exploiting the nature of CNN sampling position, we propose a novel cross-modal fusion design (MGS) for salient object detection, where we use a subsidiary modality, i.e., RGB/depth, to guide the main modality streaming, i.e., depth/RGB.
- For RGB-only input, we suggest using an off-the-shelf depth prediction model to mimic the multi-modality input. Our MGSnet enables dramatical performance gain on benchmark datasets and achieves state-of-the-art performance among RGB SOD models.
- The proposed MGS can also be embedded in RGBD two-stream network with the advantage of cross-modality cues while being lightweight.

2. Related Work

**RGB SOD:** In the past decade, the development of GPU and CNN contributes to the advances of RGB SOD. One core problem is understanding the geometric information from the image. Fully Convolutional Network (FCN) [21] is a pioneering work in leveraging spatial information in CNN. Most recent researches dominating RGB SOD are FCN-based, such as [33] which designs a single stream encoder-decoder system, [15] which adopts a multi-scale network on input, and most currently [5, 19, 32, 35] which fuse multi-level feature map. Some branch designs also have achieved impressive results such as C2S-Net [17] which bridges contour knowledge for SOD. By inserting additional transformation parameters in networks, it contributes to the model performance. Nevertheless, the inference time and computational cost become more significant.

**RGBD SOD:** The complementary depth map may provide extra clues on the geometry. How to efficiently joint RGB and depth modality is the key challenge for RGBD SOD. One possible solution is to treat the depth map as...
an additional channel and adapt a single-stream system as shown in DANet [36]. It further designs a verification process with a depth-enhanced dual attention module. An alternative is to realize multi-stream networks followed by a feature fusion mechanism. PDNet [37] designs a depth-enhanced stream to extract geometric features and further fuses with the RGB features. D3net [7] adopts separate networks to respectively extract features from RGB, depth map, and RGBD four-channel input. A late fusion is further realized. HDFnet [23] adopts two streaming networks for both RGB image and depth map. These features are further fused to generate region-aware dynamic filters. JL-DCF [9] proposes joint learning from cross-modal information through a Siamese network. Generally, RGBD networks achieve superior performance compared to RGB as shown in Figure 1. However, these methods rely on the quality and accessibility of the depth map. A high-quality depth map requires expensive depth sensors and is still sparse compared to an RGB image as suggested in [9, 7]. To this end, DCF [11] proposes to calibrate the raw depth to improve the quality. Nevertheless, the high computational cost due to the two-streaming network requires more development.

Some recent researches [26, 12, 34] propose to learn from RGBD images and tests on RGB. This design enables an RGB CNN to achieve a comparable result with RGBD SOD during testing. Different from it, we propose to firstly discover the hidden geometric modality behind RGB images by simply using an off-the-shelf depth prediction method. With the estimated depth, we further propose a Modality-Guided Subnetwork mechanism to enhance the master RGB network understanding of the contour problem. Our proposed MGSnet achieves state-of-the-art performance with real-time inference speed compared to other RGB models. It can also be embedded in RGBD two-stream models to enable further progress with raw depth.

3. Modality-Guided Subnetwork

3.1. Overview

In Figure 2 (c), our network only takes RGB as input that then estimates the pseudo-depth. Our MGSnet only takes the pseudo-depth to deform the RGB streaming. In other words, only the RGB modality is fed through Conv_4.

Note that our model is not limited by the nature of the modality. It can be a depth-guided RGB convolution as well as an RGB-guided depth convolution. Figure 3 presents our model embedded on an RGBD two-streaming network and Figure 4 illustrates the idea of modality-guided sampling position. We learn the offset from both semantic RGB and depth features to create a cross supervision mechanism.

For simplicity, we present in the following section a depth-guided subnetwork for RGB features. It contains three parts: a master RGB streaming network, an off-the-shelf prediction model to estimate a pseudo-depth map if not available, and a depth-guided subnetwork design. For simplicity, VGG-16 [28] architecture is adopted as our basic convolutional network to extract RGB features for its wide application in SOD. We use RFB [20] on the steamer layers ($f_3, f_4, f_5$) which contains high level features for SOD as suggested in [26, 23, 9]. We further embed our subnetwork to enhance the edge understanding of the encoder output. We take the same decoder as proposed in [26] and a simple binary cross-entropy (BCE) as the loss.

3.2. Depth-guided Subnetwork

To proceed with the geometric prior, the depth map $D$ and the RGB feature map (output of Conv_4) are fed together to our model. We use $f_i \in \mathbb{R}^{b \times 512 \times h \times w}$ to denote the input RGB feature. The depth prior and RGB feature maps are articulated through an adaptive convolution to compute depth-aware RGB feature maps as output. The last is added to the master RGB stream to form the final feature map.

The subnetwork contains three convolutions of different filter sizes: $1 \times 1, 3 \times 3$, and $1 \times 1$. It shares the same architecture of plain baseline of ResNet [10] that the $1 \times 1$ layers are used for reducing (512→64) and then increasing dimensions (64→512), allowing the $3 \times 3$ layer with smaller input/output dimensions. We denote $D$ and $U$ for the first and the last $1 \times 1$ convolution, which stands for down-sample and up-sample, respectively. This design can significantly reduce the learning parameters, which contributes to the lightweight design of our subnetwork. Different from ResNet that uses the three layers as a bottleneck, we use them as the residual-addition branch which serves as complementary information to the plain network.

A standard convolution is formulated as:

$$y(p) = \sum_{p_n \in R(p)} w(p_n) \cdot x(p + p_n).$$

with $R(p)$ the sampling grid for point $p$. Physically it represents a local neighborhood on input feature map, which conventionally has regular shape with certain dilation $\Delta d$, such that:

$$R(p) = a\mathbf{u} + b\mathbf{v}$$
where \((\tilde{u}, \tilde{v})\) is the pixel coordinate system of input feature map and \((a, b) \in (\Delta d \cdot \{-1, 0, 1\})^2\) for a \(3 \times 3\) convolution or \((a, b) \in \{(0)\}^2\) for \(1 \times 1\) convolution.

We further replace the conventional \(3 \times 3\) convolution by deformable convolution (DeformConv) [4], where the kernels are generated with different sampling distributions which is adapted to depth modality. Mathematically, we have:

\[
y(p) = \sum_{p_n \in \mathcal{R}(p)} w(p_n) \cdot x(p + p_n + \Delta p_n) \tag{3}
\]

The convolution may be operated on the irregular positions \(p_n + \Delta p_n\) as the offset \(\Delta p_n\) may be fractional. To address the issue, we use the bilinear interpolation which is the same as that proposed in [4]. The adapted convolution is denoted as \(\mathcal{A}\).

Thanks to the depth input of the subnetwork, the scale and geometric transformation of objects on the RGB feature map can be directly analyzed with the adapted offsets. This process is expressed as:

\[
\Delta p_n = \eta(D) \tag{4}
\]

We present two types of offset generators according to different plain networks. More details are discussed in the following section. The newly defined sampling position becomes depth-aware and helps to better articulate the RGB feature and geometric information. Finally, the output of MGS is added to the master stream, which serves as complementary depth-aware guidance on RGB features.

The entire process to compute the modality-guided feature \(f_M\) can be formulated as follows:

\[
f_M = MGS(f_s, D) = \mathcal{U}(\mathcal{A}(\mathcal{D}(f_s), \eta(D))) \tag{5}
\]

The output of RGB encoder can be formulated as:

\[
out = f_s + \lambda f_M \tag{6}
\]

where \(\lambda\) is the weight parameter.

### 3.3. Offset generator

We use another modal prior to deform the main stream convolution. When the offset exceeds the input size, the output will be computed as if the zeros padding is applied. For RGB input, the pseudo-depth is used to deform the RGB sampling position. The offset is generated through Z-ACN [31]. It firstly back-projects the 2D conventional grid to form a 3D point cloud according to the depth. Based on the point cloud, it extracts a depth-aware 3D plan and further creates a depth-aware 3D regular grid. Then it projects the 3D regular grid to the image plan to form the deformable sampling position. More details can be found in Z-ACN [31] paper. Different to DeformConv [4] that learns offset from the RGB feature map to deform RGB sampling position, Z-ACN computes offset according to low-level geometric constraint (one-channel depth) and does not require gradient descent, thus perfectly matches our light-weight subnetwork design. The computed offset allows the RGB convolution to be scale and rotation independent. We verify through experiments the superior performance of our model in the ablation study.

For RGBD input, current Sconv [2] suggests learning the RGB offset from a semantic depth feature map. We share the same motivation as Sconv. However, Sconv firstly projects the depth into a high-dimensional feature space and secondly learns a depth-aware offset and mask. Unlike Sconv, we learn the offset from the encoder or high-level features to avoid the additional projection. In other words, in our case, the offset generator \(\eta\) is realized through a simple \(3 \times 3\) convolution to minimize the computational cost. Furthermore, we adapt to different modalities as input, i.e., it learns offset from both RGB and depth, while Sconv only learns from depth.

### 3.4. Understand adaptive sampling position

Our model aims to compensate for the single modality streaming. As shown in Figure 4, while extracting features from RGB images, the conventional sampling position is
limited by the lack of capability to include geometry due to
the fixed shape. We propose to use the depth prior to accu-
rately locate the sampling position. For RGB input without
depth prior, we suggest mimicking the depth map by using
a monocular depth estimation model. Some pseudo-depth im-
ages may be inaccurate due to the domain gap between SOD
and monocular depth estimation. In such a case, the offset
will converge to 0 so that the deformation becomes minimal
and local. The contribution of the depth-aware RGB feature
is further regularized by the weight parameter $\lambda$ of Eq. 6.
In Fig. 5, we show that our method is robust to non-optical
depth through several examples.

While extracting features from raw depth, conventional
sampling positions may produce sub-optimal results due to
some inaccurate measurements. The raw depth maps for
SOD are obtained by camera measurements such as Kinect
and Light Field cameras, or estimated by classic computer
vision algorithms as [29, 18]. Thus, the raw depth images
may contain noise and ambiguity. We can visualize several
low-quality samples on the third row of Figure 5. To this
end, we propose to use the RGB image to deform the depth
sampling position. In such a case, the RGB-guided sam-
pling position can make up for the measurement error on
geometry.

4. Experiments

4.1. Benchmark Dataset

To verify the effectiveness of our method, we conduct ex-
periments on seven following benchmark RGBD datasets.
DES [3] : includes 135 images about indoor scenes cap-
tured by Kinect camera. LFSD [16] : contains 100 images
collected on the light field with an embedded depth map and
human-labeled ground truths. NLPR [24] : contains 1000
natural images captured by Kinect under different illumi-
ation conditions. NJUD [13] : contains 1,985 stereo im-
age pairs from different sources such as the Internet, 3D
movies, and photographs taken by a Fuji W3 stereo cam-
era and with estimated depth by using optical flow method
[29]. SSD [38] : contains 80 images picked up from stereo
movies with estimated depth from flow map [29]. STEREO
[22] : includes 1000 stereoscopic images downloaded from
the Internet where the depth map is estimated by using SIFT
captured by Lytro camera in real-life scenes.

4.2. Experimental Settings

Our model is implemented basing on the Pytorch tool-
box and trained with a GTX 3090Ti GPU. We adopt several
generally-recognized metrics for quantitative evaluation: F-
measure is a region-based similarity metric that takes into
account both Precision (Pre) and Recall (Rec). Mathemati-
cally, we have: $F_\beta = \frac{(1+\beta^2) \cdot Pre \cdot Rec}{\beta^2 \cdot Pre + Rec}$. The value of $\beta^2$ is set
to be 0.3 as suggested in [1] to emphasize the precision.
In this paper, we report the maximum F-measure ($F_\beta$)
score across the binary maps of different thresholds, the
mean F-measure ($F_\beta^{mean}$) score across an adaptive thresh-
hold and the weighted F-measure ($F_\beta^w$) which focuses more
on the weighted precision and weighted recall. Mean Ab-
solute Error (MAE) studies the approximation degree be-
tween the saliency map and ground-truth map on the pixel
level. S-measure ($S_m$) evaluates the similarities between
object-aware ($S_o$) and region-aware ($S_r$) structure between
the saliency map and ground-truth map. Mathematically, we
have: $S_m = \alpha \cdot S_o + (1 - \alpha) \cdot S_r$, where $\alpha$ is set to be 0.5.
E-measure ($E_m$) studies both image level statistics and lo-
cal pixel matching information. Mathematically, we have:
$E_m = \frac{1}{m \times n} \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} \phi_{FM}(i, j)$, where $\phi_{FM}(i, j)$ stands for
the enhanced-alignment matrix as presented in [6].

4.3. Performance Comparison with RGB Input

We firstly compare with RGB models, including R3Net
[5], PoolNet [19], CPD [32], AFnet [8]. All saliency maps
are directly provided by authors or computed by authorized
codes. For fair comparisons, we adopt the same training
set as suggested in [26], which contains 1485 samples from
NJUD, 700 samples from NLPR, and 800 samples from the
DUT-RGBD dataset. The remaining images of all listed
datasets are used for testing. The quantitative comparison is
presented in Table 1. Our model is trained with 50 epochs
with 256 × 256 input image size.

For the RGB model, we can conclude from Table 1 that
the improvement on the saliency map is attributed to differ-
ent learning modules, which results in high computational
cost (size). Different from traditional RGB models which
do not exploit the depth information, we propose to take
full advantage of the pseudo-geometry estimated with an
existing monocular depth estimation method.

We re-train two RGB-D SOD network (HDFnet [23],
CoNet [12]) with the additional estimated pseudo-depth.
We observe a significant performance gap between the re-
cent RGB-D models and the previous RGB models. The
main reason is the quality of depth estimation: the do-
main gap between the depth estimation dataset and the SOD
dataset leads to some failure depth maps. This can be not-
iced in the poor performance of HDFnet that extracts fea-
tures from both RGB and depth images. CoNet, however,
is more robust to the depth quality since the depth map is
only used to supervise the feature extraction on RGB im-
geagues. Our model shares the same motivation as CoNet to
use depth prior to guide SOD but in a completely different
manner. In our model, we directly learn a geometric-aware
offset from the depth map to the sampling position on the
RGB image. Our model achieves consistent superior per-
formance compared with other models.
4.4. Performance Comparison with RGB-D Input

We also compare with state-of-the-art RGBD models with raw depth input in the Table 2, including CoNet [12], A2dele [26], DANet [36], cmMS [14], HDFnet [23], and DSA2F [30]. For fair comparisons, all saliency maps and the FPS are directly provided by authors or computed by authorized codes. Note that the FPS depends on the GPU for inference. Thus, only the FPS of HDFnet is tested on the same GPU as ours.

While depth is only used as supervision during training and only RGB image is required during testing, our model surpasses existing efficient A2dele significantly on performance with only an + around 5Mb model size. Compared to CoNet, the model size is minimized by 63% and achieves a comparable result. As presented in Figure 4, our proposed module can take advantage of cross-modality cues while being lightweight. Thus, we further incorporate with the HDFnet [23] to show the performance gain by integrating our approach. It achieves the state-of-the-art (SOTA) performance on VGG16 based models (HDF + Ours). To better demonstrate the superiority of the proposed method, we also use a larger backbone (VGG19) to compare with the plain version HDFnet and the SOTA method DSA2F. Note that DSA2F uses neural architecture search to automate the model architecture while ours is hand-designed. Our model enables significant gains on the plain version with minimal cost (+ around 1 Mb on model size) and achieves comparable results with the DSA2F.

4.5. Qualitative Evaluation

We present the qualitative result with some challenging cases in Figure 5: low density (1st column), similar visual appearance between foreground and background (2nd – 5th columns), small objects (6th columns), far objects (7th – 9th columns), human in scene (10th columns), and similar and low contrast on depth map (11th – 13th columns). It can be seen that our MGSnet yields the results closer to the ground truth mask in various challenging scenarios, especially for the last three columns with low-quality depth clues. Different from two-stream networks that tend to treat sub-optimal depth equally as RGB input, MGSnet extracts features from RGB images while the depth map serves only as complementary guidance, thus becoming robust to depth bias. By analyzing the response on HDFnet (sixth row) and HDFnet with embedded MGS (seventh row), we observe that our approach enables the plain network better discrimination of salient objects from the background.

5. Ablation Study

Effect of Modality-Guided Sampling Position: Our modality-guided sampling position aims to incorporate multi-modal information through the basic function of CNN - the sampling position of convolution. This pattern is integrated in Eq. 3 and Eq. 4. To verify the effectiveness of the proposed modality-guided sampling position, a series of experiments with different learning strategies are realized.

(1) - (4) are experiments on RGB model: (1) RGB Baseline. (2) Self-guided deformable sampling position. We learn the offset from the RGB feature map. (3) RGB pseudo-depth early fusion. We form a four-channel input with pseudo depth. (4) Depth-guided deformable position. We compute an offset from pseudo-depth using ZACN to guide RGB streaming. (5) - (7) are experiments on RGBD model: (5) Baseline. We use the same architecture as HDFnet. (6) Self-guided deformable sampling position. The offset applied to RGB streaming is learned from the RGB feature. Idem for depth streaming. (7) Cross modality-guided deformable position. We learn an offset from depth to guide RGB streaming, and vice versa.

Table 3 (1) and (3) compare the performance of the baseline RGB three-channel input and mimicked RGBD four-channel input with pseudo-depth, respectively. The mim-

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Size ↓</th>
<th>DES</th>
<th>NLPR</th>
<th>NJUD</th>
<th>STEREO</th>
<th>DUT-RGBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>R3Net18</td>
<td>MAE ↓</td>
<td>0.66</td>
<td>0.728</td>
<td>0.693</td>
<td>0.101</td>
<td>0.649</td>
<td>0.611</td>
</tr>
<tr>
<td>PoolNet19</td>
<td>MAE ↓</td>
<td>0.031</td>
<td>0.852</td>
<td>0.814</td>
<td>0.046</td>
<td>0.791</td>
<td>0.771</td>
</tr>
<tr>
<td>CPD19</td>
<td>MAE ↓</td>
<td>0.112</td>
<td>0.828</td>
<td>0.841</td>
<td>0.037</td>
<td>0.840</td>
<td>0.829</td>
</tr>
<tr>
<td>AFNet19</td>
<td>MAE ↓</td>
<td>0.144</td>
<td>0.840</td>
<td>0.816</td>
<td>0.043</td>
<td>0.807</td>
<td>0.796</td>
</tr>
<tr>
<td>EGNNet19</td>
<td>MAE ↓</td>
<td>0.412</td>
<td>0.831</td>
<td>0.797</td>
<td>0.047</td>
<td>0.800</td>
<td>0.774</td>
</tr>
<tr>
<td>+ Pseudo Depth (86 Mb extra model size)</td>
<td>MAE ↓</td>
<td>0.721</td>
<td>0.664</td>
<td>0.062</td>
<td>0.758</td>
<td>0.741</td>
<td>0.124</td>
</tr>
<tr>
<td>CoNet20</td>
<td>MAE ↓</td>
<td>0.171</td>
<td>0.820</td>
<td>0.808</td>
<td>0.049</td>
<td>0.744</td>
<td>0.835</td>
</tr>
<tr>
<td>Ours</td>
<td>MAE ↓</td>
<td>0.171</td>
<td>0.820</td>
<td>0.808</td>
<td>0.049</td>
<td>0.744</td>
<td>0.835</td>
</tr>
<tr>
<td>Backbone</td>
<td>Extract RGB feature</td>
<td>Extract RGD feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>---------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CoNet&lt;sub&gt;20&lt;/sub&gt;</td>
<td>DANet&lt;sub&gt;20&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A2dele&lt;sub&gt;20&lt;/sub&gt;</td>
<td>cmMS&lt;sub&gt;20&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>HDFnet&lt;sub&gt;20&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- VGG16 –</td>
<td>- VGG16 –</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- RGB –</td>
<td>- RGBD –</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- VGG16 –</td>
<td>- VGG19 –</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size ↓</td>
<td>167</td>
<td>430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FPS ↑</td>
<td>57</td>
<td>177</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE ↓</td>
<td>.027</td>
<td>.023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{F}_{\text{mean}}^{\beta}$ ↑</td>
<td>.826</td>
<td>.887</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_m$ ↑</td>
<td>.910</td>
<td>.904</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_m$ ↑</td>
<td>.945</td>
<td>.967</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE ↓</td>
<td>.031</td>
<td>.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{F}_{\text{mean}}^{\beta}$ ↑</td>
<td>.848</td>
<td>.871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_m$ ↑</td>
<td>.908</td>
<td>.915</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_m$ ↑</td>
<td>.934</td>
<td>.949</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE ↓</td>
<td>.047</td>
<td>.045</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{F}_{\text{mean}}^{\beta}$ ↑</td>
<td>.872</td>
<td>.871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_m$ ↑</td>
<td>.895</td>
<td>.899</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_m$ ↑</td>
<td>.911</td>
<td>.922</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE ↓</td>
<td>.037</td>
<td>.047</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{F}_{\text{mean}}^{\beta}$ ↑</td>
<td>.885</td>
<td>.858</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_m$ ↑</td>
<td>.908</td>
<td>.901</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_m$ ↑</td>
<td>.928</td>
<td>.914</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Quantitative comparisons of with recent RGBD models. ↑ & ↓ denote larger and smaller is better, respectively. MGS can also be embedded to the HDFnet [23] to enable further progress. The scores/numbers better than ours are underlined (extracting RGB feature, extracting RGBD feature with VGG16, and extracting RGBD feature with VGG19 models are labeled separately). More details on all RGB-D dataset can be found in the Supplementary Material.

Figure 5: Visual comparison between our proposed MGSnet and the state-of-the-art RGB/RGBD methods. * denotes that the ground truth depth is used during testing. We also embed MGS on the HDFnet [23] to enable further improvement, denoted as *HDF + Ours. More examples can be found in the Supplementary Material.
Table 3: Ablation study of modality-guided sampling position

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DES</th>
<th>LFSD</th>
<th>NLPR</th>
<th>NJUD</th>
<th>SSD</th>
<th>STEREO</th>
</tr>
</thead>
<tbody>
<tr>
<td># Metric</td>
<td>MAE</td>
<td>F_m</td>
<td>MAE</td>
<td>F_m</td>
<td>MAE</td>
<td>F_m</td>
</tr>
<tr>
<td>1 RGB Baseline</td>
<td>.036</td>
<td>.880</td>
<td>.104</td>
<td>.835</td>
<td>.037</td>
<td>.883</td>
</tr>
<tr>
<td>2 RGB + Self Deform</td>
<td>.042</td>
<td>.860</td>
<td>.110</td>
<td>.798</td>
<td>.031</td>
<td>.885</td>
</tr>
<tr>
<td>3 RGB pseudo D Fusion</td>
<td>.032</td>
<td>.888</td>
<td>.093</td>
<td>.819</td>
<td>.029</td>
<td>.893</td>
</tr>
<tr>
<td>4 RGB + Depth-Deform</td>
<td><strong>.028</strong></td>
<td><strong>.899</strong></td>
<td><strong>.078</strong></td>
<td><strong>.849</strong></td>
<td><strong>.025</strong></td>
<td><strong>.905</strong></td>
</tr>
<tr>
<td>5 RGB-D Baseline (B)</td>
<td>.020</td>
<td>.933</td>
<td>.089</td>
<td>.856</td>
<td>.028</td>
<td>.921</td>
</tr>
<tr>
<td>6 B + Self Deform</td>
<td>.019</td>
<td>.936</td>
<td>.089</td>
<td>.853</td>
<td>.026</td>
<td>.916</td>
</tr>
<tr>
<td>7 B + Cross-Modal Deform</td>
<td><strong>.019</strong></td>
<td><strong>.936</strong></td>
<td><strong>.079</strong></td>
<td><strong>.871</strong></td>
<td><strong>.025</strong></td>
<td><strong>.921</strong></td>
</tr>
</tbody>
</table>

Figure 6: Visual analysis of embedded depth with MGSnet.

icked multi-modality early fusion achieves better performance, indicating that the pseudo-depth provides additional semantic. However, by comparing (3) and (4), we observe that the proposed depth-guided deformable sampling position can better use the complementary information to supervise RGB streaming, compared with early fusion. By comparing (2) and (4), we show that the depth-guided deformable position is more accurate on saliency compared to that of the self/RGB-guided. This verifies the assumption that depth cues can help the RGB model to better distinguish the foreground and background. Note that in (4) we only extract features from RGB images. The additional awareness of the geometry is only treated as a 2D offset to better locate the sampling position. This new integration design contributes to the model performance with minimal cost. For better understanding, the qualitative result presented in Figure 6 shows that our approach provides more accurate saliency maps with better contrast. On the RGBD model (5-7), we also observe the superior performance with the cross-modality deformable sampling position achieves as it directly compensates for the single modal streaming.

Performance with different depth qualities: We also conduct an experiment to show the impact of depth quality. We choose the HDFnet [23] as the baseline and further embed it with our method. We present the average metric on all testing datasets in Table 4 with pseudo-depth (estimated) and raw depth from the RGBD dataset. Results obtained with pseudo-depth are denoted with *.

It shows that the quality of depth has an important influence on performance. Features extracted from raw depth describe better the salient object and were in line with our expectations. However, in both cases, our MGS can significantly enable progress compared to the plain networks. For pseudo-depth, the contribution of our MGS is more significant, which can be explained by the effectiveness of our RGB-guided sampling position for depth streaming. It can efficiently help to alleviate depth errors.

6. Conclusions

In this paper, we propose a modality-guided module (MGSnet) for RGB-D salient object detection models. The depth channel can either be the input or be estimated using a state-of-the-art monodepth network. Our model adaptively transforms convolutions that their size and shape are built by fusing information from one modality to another (e.g., depth to RGB and/or RGB to depth), thus enabling a cross-modality interaction. Extensive experiments against RGB baselines demonstrate the performance gains of the proposed module, and the addition of the proposed module to existing RGB-D models further improved results.

Acknowledgements

We gratefully acknowledge Zhuyun Zhou for her support and proofreading. We also thank Jilai Zheng, Zhongpai Gao, and Yonglin Zhang for the discussion. This research is supported by the French National Research Agency through ANR CLARA (ANR-18-CE33-0004) and financed by the French Conseil Régional de Bourgogne-Franche-Comté.
References


[14] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In European Conference on Computer Vision (ECCV), 2020. 6


[23] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In European Conference on Computer Vision (ECCV), 2020. 1, 2, 3, 5, 6, 7, 8


[27] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset
transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 6


[34] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020. 3


[38] Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017. 5