Robust RGB-D Fusion for Saliency Detection

Zongwei Wu, Shriarulmozhivarman Gobichettipalayam, Brahim Tamadazte, Guillaume Allibert, Danda Pani Paudel, Cédric Demonceaux

▶ **To cite this version:**

Zongwei Wu, Shriarulmozhivarman Gobichettipalayam, Brahim Tamadazte, Guillaume Allibert, Danda Pani Paudel, et al.. Robust RGB-D Fusion for Saliency Detection. 10th International Conference on 3D Vision, Sep 2022, Prague, Czech Republic. hal-03746242v1

HAL Id: hal-03746242

https://hal.science/hal-03746242v1

Submitted on 5 Aug 2022 (v1), last revised 30 Aug 2022 (v2)

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.

Robust RGB-D Fusion for Saliency Detection

Zongwei Wu[1,2] Shriarulmozhivarman Gobichettipalayam[1] Brahim Tamadazte[3]
Guillaume Allibert[4] Danda Pani Paudel[2] Cédric Demonceaux[1]
[1] ImViA, Université Bourgogne Franche-Comté, France [2] Computer Vision Laboratory, ETH Zurich, Switzerland
[3] Sorbonne Université, CNRS, ISIR, France [4] Université Côte d'Azur, CNRS, I3S, France

Abstract

Efficiently exploiting multi-modal inputs for accurate RGB-D saliency detection is a topic of high interest. Most existing works leverage cross-modal interactions to fuse the two streams of RGB-D for intermediate features' enhancement. In this process, a practical aspect of the low quality of the available depths has not been fully considered yet. In this work, we aim for RGB-D saliency detection that is robust to the low-quality depths which primarily appear in two forms: inaccuracy due to noise and the misalignment to RGB. To this end, we propose a robust RGB-D fusion method that benefits from (1) layer-wise, and (2) trident spatial, attention mechanisms. On the one hand, layer-wise attention (LWA) learns the trade-off between early and late fusion of RGB and depth features, depending upon the depth accuracy. On the other hand, trident spatial attention (TSA) aggregates the features from a wider spatial context to address the depth misalignment problem. The proposed LWA and TSA mechanisms allow us to efficiently exploit the multi-modal inputs for saliency detection while being robust against low-quality depths. Our experiments on five benchmark datasets demonstrate that the proposed fusion method performs consistently better than the state-of-the-art fusion alternatives. The source code is publicly available at: https://github.com/Zongwei97/RFnet.

1. Introduction

Saliency detection aims to segment image contents that visually attract human attention the most. Existing RGB-based saliency detection methods [21, 55, 49, 57] achieve promising results in generic settings. However, in cluttered and visually similar backgrounds, they often fail to perform accurate detection. Therefore, many recent works [57, 34, 13, 5] exploit image depths as additional geometric cues, in the form of RGB-D inputs, to improve the saliency detection performance in difficult scenarios.

Given accurate and well-aligned depths, existing RGB-D methods perform well even in difficult scenarios. Unfortunately, this is not often the case in practice. Sometimes, only low-quality depths can be acquired, depending upon the scene and the source of depths. For example, depths from multi-view stereo cameras are often noisy [50, 2] and asynchronous depth cameras are spatially misaligned [31], as shown in Figure 1. Other environmental factors such as object distance, texture, or even lighting conditions during the acquisition can also degrade the depth quality [5, 47, 12, 56]. Therefore, a method that can still exploit the geometric cues, while being robust to the depth quality discrepancy is highly desired.

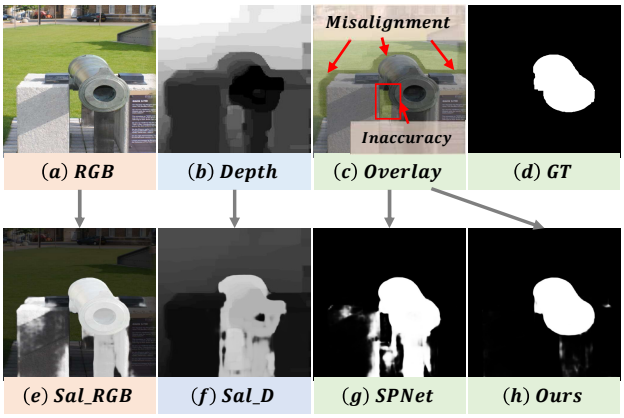

Figure 1. **Motivation**. (a) and (b) are the paired RGB-D inputs. (e) and (f) are the associated saliency maps generated from the single-modal input which are sub-optimal. (c) is the RGB-D overlay. The state-of-the-art model [62] fails to reason about accurate saliency map under inferior conditions, i.e., inaccurate depth measurement and misalignment between both modalities (g). To address this issue, we propose a robust RGB-D fusion to explicitly model the depth noise for saliency detection, yielding results (h) closer to the ground-truth mask (d).

We observe that most existing methods perform unsatisfactorily on datasets with low-quality depths. This is primarily because of the commonly used fusion technique [6, 30, 52, 26, 62, 45] that merges the parallel streams of RGB and depth with equal importance while being agnostic to misalignment. Less accurate depths are expected to play a smaller role than their counterpart. On the other hand, the

possibility of misalignment between RGB and depth needs to be considered during the fusion process.

In this work, we propose a robust RGB-D fusion method that addresses the aforementioned problems of inaccurate and misaligned depths. The proposed method uses a layer-wise attention (LWA) mechanism to enable the depth quality aware fusion of RGB and depth features. Our LWA attention learns the trade-off between early and late fusions, depending upon the provided depth quality. More precisely, LWA encourages the early fusion of the depth features for high-quality depth inputs, and vice versa. Such fusion avoids the negative influence of the spurious depths while being opportunistic when high-quality depths are provided. In other words, the good-quality depth should play an important role in early layers thanks to its rich and exploitable low-level geometric cues, while low-quality depth should be more activated at semantic levels.

To address the problem of misaligned depths, we introduce the trident spatial attention (TSA) that aggregates features from a wider spatial context. The introduced TSA is used to replace vanilla spatial attention, enabling the aligned aggregation of the misaligned features. In particular, our TSA requires only minor additional parameters and computation, while being sufficient to address the problem of misalignment. Note that the misalignment problem often exists only locally therefore the global context (at the cost of additional computation) may not be necessary. Such an example is shown in Figure 1(c). We improve the vanilla spatial attention with different scales of receptive fields, yielding a simple yet efficient manner to replace the pixel-wise correspondence with region-wise correlation. Finally, the new spatial attention is adaptively merged with channel attention to form our hybrid fusion module.

In summary, our major contributions are listed below:

- We study the problem of RGB-D fusion in a real-world setting, highlighting two major issues, inaccurate and misaligned depths, for accurate saliency detection.
- We introduce a novel layer-wise attention (LWA) to automatically adjust the depth contribution through different layers and to learn the best trade-off between early and late fusion with respect to the depth quality.
- We design a trident spatial attention (TSA) to better leverage the misaligned depth information by aggregating the features from a wider spatial context.
- Extensive comparisons on five benchmark datasets validate that our fusion performs consistently better than state-of-the-art alternatives while being very efficient.

2. Related Work

There are extensive surveys of salient object detection [42, 1, 61] and on attention modules [40, 16] in the literature. In the following, we briefly review related works.

RGB-D Fusion for Saliency Detection: In the literature, we can divide current models into two types of architectures: single-stream and multi-stream schemes. The main difference is in the number of encoders. Single-stream networks are commonly lighter compared to multi-stream works. In [9, 60], the authors proposed to concatenate RGB-D images from the input side and then feed them into a single encoder-decoder architecture. From another perspective, [34] introduces a depth distiller to enable cross-modal knowledge distillation, leading to a lightweight inference strategy with RGB-only input. Other works [58, 13, 39] propose to directly integrate low-level geometric cues in the RGB stream to strengthen the RGB features. Despite the proven result in previous works, single-stream models fail to explicitly analyze cross-modal correlation in complex scenarios, which is the main performance bottleneck.

Recently, multi-stream architectures have drawn increasing research interests. Several works [30, 6, 47, 62, 53, 26, 7] propose to explicitly model RGB and depth cues through two parallel encoders and then aggregate multi-modal features through multi-scale fusion schemes, leading to better performance compared to their counterpart. In the literature, we can group existing works into three categories based on the fusion schemes: 1) depth-guided fusion, 2) discrepant fusion, and 3) multi-scale fusion. Depth enhanced fusion models [6, 30, 56] often adopts an asymmetric fusion scheme that the depth features are fused into RGB features at each level to improve boundary awareness. However, these models are sensitive to depth noise and the performance is significantly degraded when depth maps are under inferior conditions. Other works [30, 54, 47, 52, 12] propose to merge multi-modal cues through a discrepant design. In [52], the authors adopt different fusion designs for low-level and high-level features, i.e., RGB to calibrate depth in earlier layers and depth to calibrate RGB in deeper layers. [47] adopts lightweight spatial attention [46, 48] only at semantic level. [54, 12] only fuse features at semantic levels, i.e., outputs from the last three layers. Different from discrepant and asymmetric designs, a number of works [57, 62, 53, 7] realize bi-directional cross-modal interaction at each scale of the neural network. This fusion design, also known as middle fusion, has shown plausible performance in saliency benchmarks. Nevertheless, we observe that most existing works treat RGB and depth equally to form the shared features, paying little attention to explicitly modeling the measurement bias and alignment issue. [56] has introduced a weighting strategy to deal with the measurement bias. However, their weighting scheme assumes the perfect alignment between multi-modal features. Different from previous works, we estimate the depth quality index by leveraging contextualized awareness. We show through empirical comparison that our approach can better model the depth quality to adjust the contribution.

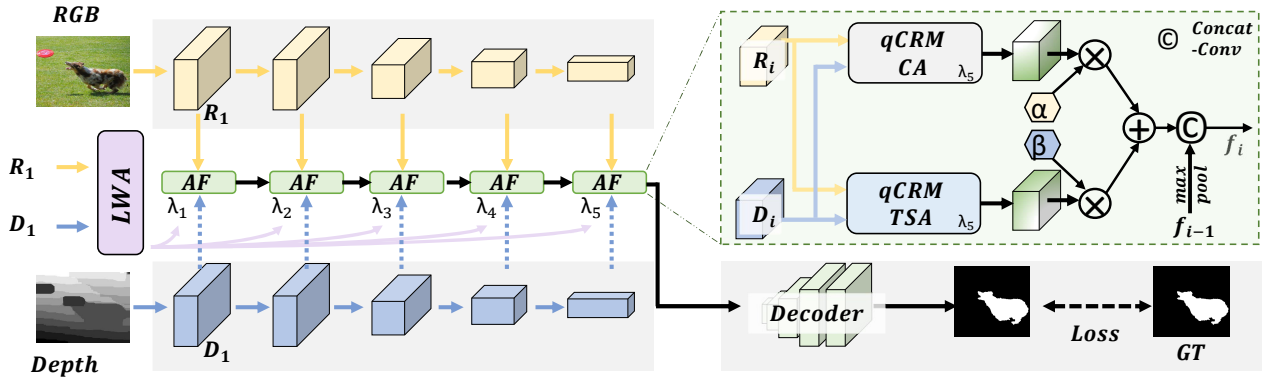

Figure 2. **Architecture**. Our proposed network consists of a layer-wise attention (LWA, see Section 3.1) and an adaptive Attention Fusion (AF, see Section 3.2). LWA aims to find the best trade-off for early and late fusion depending on the depth quality, while AF leverages cross-modal cues to compute the shared representation with channel attention and improved spatial attention (TSA). CRM is from [12].

Attention for Cross-Modal Interaction Self-attention modules [41, 44, 43, 8, 35, 25] have been proven to be efficient for visual tasks. Inspired by their success, several RGB-D saliency works [6, 57, 12, 22, 52] leverage self-attention as an augmentation to better preserve, calibrate, and fuse multi-modal features. [57, 12] explicitly leverages the attention along the channel direction to calibrate each modality. [22] introduces a mutual and non-local strategy to learn the spatial cues from one modality and apply it to the other. Several recent works [26, 23, 7] further explore the long-range dependencies with transformer attention [41].

Despite the popularity of contextualized attention, we observe that these modules often require a significant computational cost. Therefore, fusion with transformer attention is often realized with a small resolution feature map, i.e., at deeper layers of encoders [22, 26, 7]. To benefit from the spatial cues at each stage, a number of works [6, 52, 7] adopt the hybrid models with vanilla spatial and channel attention from [44] to aggregate features at each stage. However, vanilla spatial attention is agnostic of feature misalignment. Moreover, these hybrids treat spatial and channel attention equally, failing to be adjusted with respect to the network depth. Unlike previous works, we propose a simple yet efficient trident spatial attention that can better model contextualized awareness than its counterpart. Furthermore, we integrate our spatial attention with channel attention in a parallel scheme, yielding a more robust fusion strategy with adaptive weights.

3. Method

Before introducing the details, we highlight our technical motivation for better understanding the novelty of the proposed technique. D3Net [5] is one of the pioneering works that explicitly model both modality-specific and fused saliency maps. The fusion design is realized at the output/saliency level. Differently, in our network, the RGB-D features are merged at the encoder stage. Several recent works [19, 18] fuse RGB-D features during encoder with the help of spatial attention. [19] uses RGB cues to improve depth, while [18] is bi-directional. From another perspective, DFMNet [56] learns a weight to adjust the depth contribution. However, these methods focus more on the depth quality, paying little attention to the misalignment, i.e., vanilla spatial attention [44] for [19, 18] or pixel-wise add/mul for [56]. Differently, we explicitly decouple the low-quality and misalignment. We first leverage global attention to purely analyze the depth quality, based on which we introduce layer-wise attention to learn the best trade-off between early and late fusion. We show in Table 3 that our method outperforms the concurrent [56]. Moreover, we propose an improved version of spatial attention with enlarged receptive fields. Compared to vanilla spatial attention, we show in Figure 4 and Table 3 that our improved version can better leverage multi-scale cues to tackle feature misalignment and yield superior performance.

Figure 2 presents the overall framework of our Robust Fusion network (RFNet). We first extract RGB and depth features through parallel encoders. Then, these features are gradually merged through our proposed fusion module with respect to the depth noise. Specifically, to tackle the inaccurate measurement bias, we propose layer-wise attention (LWA) to control the depth contribution. To deal with feature misalignment, we propose a hybrid attention fusion (AF) module with a trident spatial attention and an adaptively merged channel attention. Details of each component are presented in the following sections.

3.1. Layer-Wise Attention

We observe that there exist several depths with unsatisfactory quality as shown in Figure 1. Inspired by this observation, we propose a depth quality indicator that aims to explicitly model the depth contribution. Our intuition is that while dealing with low-quality depth at early layers, the network should have a higher confidence value on the RGB

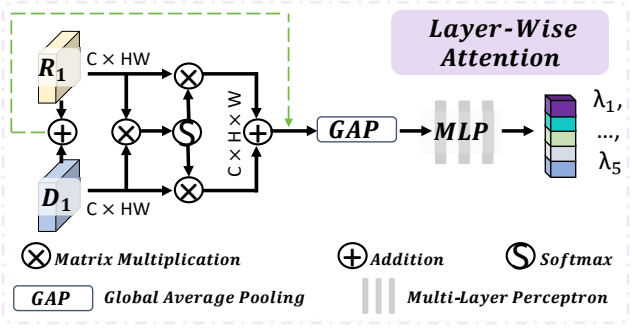

Figure 3. **Layer-Wise Attention (LWA)**. It takes paired RGB and depth low-level features as input, i.e., features from first layer R_1 and D_1, and outputs confidence values λ_i to adjust the depth contribution for the i^{th} stage fusion. Specifically, we first leverage non-local attention to enable bi-directional interaction. Then, the cross-calibrated features are merged together and fed into an MLP to model the depth contribution. The dashed shortcut (**green**) stands for the residual addition for reducing gradient vanishment.

feature instead of equally average the multi-modal cues.

As depicted in Figure 3, our layer-wise attention takes outputs from the first encoder layer as input, i.e., $R_1 \in \mathbb{R}^{C\times H\times W}$ and $D_1 \in \mathbb{R}^{C\times H\times W}$. We argue that these features contain more heterogeneous and modality-specific cues compared to semantic-level features which are homogenized. With R_1 and D_1, we first compute the similarity between the two modalities. Instead of directly realizing the pixel-wise multiplication, we leverage the contextualized awareness to avoid the feature misalignment and focus on the measurement bias. Specifically, R_1 and D_1 are firstly fed into $Conv_{1\times1}$ and flattened to form $R_1' \in \mathbb{R}^{C\times HW}$ and $D_1' \in \mathbb{R}^{C\times HW}$. These new features are then fed into the matrix multiplication. To normalize the obtained attention map, we further apply the softmax function to adjust the weight. Further, the normalized weight map is multiplied to flattened R_1 and D_1 to improve the cross-modal awareness. Finally, the retrieved RGB and depth attention maps are merged through addition. Formally, the similarity matrix can be formulated as:

$$Attention(R_1', D_1') = softmax(\frac{R_1' D_1'^T}{\sqrt{c}})(R_1' + D_1'). \tag{1}$$

Similar to self-attention works [41, 43], we add a skip connection with early fused RGB-D features to stabilize the training procedure. Once we obtain the similarity matrix, we seek to explicitly quantify the depth measurement bias. Specifically, we first extract the feature vector with the help of global average pooling (GAP) and then feed it into a multi-level perceptron (MLP) to estimate the confidence values. We particularly estimate distinct values to explicitly guide feature fusion at different scales. The adaptive weight $\lambda \in \mathbb{R}^5$ can be formulated as:

$$\lambda = MLP(GAP(Attention(R_1', D_1'))). \tag{2}$$

Finally, let R_i and D_i be the encoded RGB-D features from the i^{th} layer. Instead of equally averaging both feature maps by $R_i + D_i$ which is agnostic of input depth quality, our proposed fusion by $R_i + \lambda_i D_i$ can better merge multi-modal features with context awareness.

At first glance, our attention map is similar to non-local attention [43] which has been applied in S2MA [22] or to transformer attention [41] which has been applied in TriTrans [26]. However, our method differs from previous works in two aspects, i.e., the purpose and the model size. Compared to S2MA which uses non-local attention for cross-modal calibration, our work aims to analyze the similarity between multi-modal features and assign a confidence value to the depth cues. Compared to TriTrans which adopts multi-head transformer attention to fuse features at the deepest layer, our design is significantly lighter with only one head and is applied to low-level features with higher resolution. The concurrent work DFMnet [56] adopts Dice similarity coefficient [27] to analyze the depth quality. However, it simply multiplies RGB and depth features with the pixel-wise association, paying little attention to explicitly model measurement bias and the misalignment in a separate manner.

3.2. Adaptive Attention Fusion

Existing methods [6, 57, 12, 52, 7] often adopt attention modules, i.e., spatial attention (SA) and channel attention (CA), to enable cross-modal interaction, with few methods pay attention to inherent feature misalignment. While by design CA is more robust to this issue due to the squeezed spatial resolution, the vanilla SA has more difficulties dealing with this inferior condition since it assumes a perfect alignment between different modalities. To address this dilemma, we propose to improve the current SA with enlarged global awareness, yielding a simple yet efficient manner to replace the pixel-wise alignment with region-wise correlation. Furthermore, current works simply apply CA and SA one by another [6, 7] or equally average them to form the output [52]. These works are agnostic to the network depth that SA and CA still contribute equally at each stage. Previous work [29] has shown that layers with different depths will pay attention to different contexts. Therefore, we seek to introduce an adaptive fusion strategy with learnable weights to automatically adjust the contribution of each attention at different levels.

Formally, let an input feature map $f \in \mathbb{R}^{C\times H\times W}$. The vanilla SA firstly squeezes the channel dimension with average and max pooling across the channel, denoted as $CAP(\cdot)$ and $CMP(\cdot)$, respectively, to obtain the spatial map $f' \in \mathbb{R}^{2\times H\times W}$. Then, from f', SA learns a 2-D

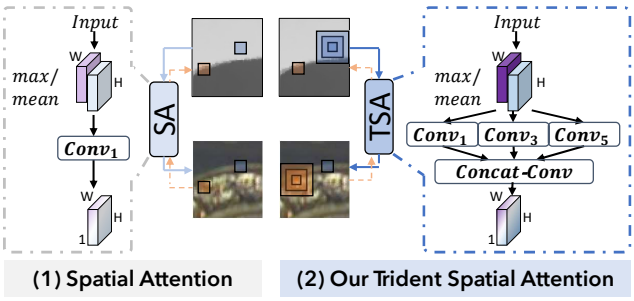

Figure 4. **Motivation of our attention fusion**: (1) Vanilla spatial attention [44, 6, 52] which is not suitable for cross-modal interaction due to feature misalignment. (2) We propose a trident spatial attention (TSA) with dilated receptive field to better leverage contextualized awareness. Better to zoom in.

weight map $SA \in \mathbb{R}^{1\times H\times W}$:

$$\begin{aligned} f' &= Concat(CAP(f), CMP(f)); \\ SA(f) &= \sigma(Conv_1(f'))), \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the Sigmoid activation, $Conv_1$ stands for the convolution with dilation 1. To improve global awareness, as shown in Figure 4 we replace the current convolution with trident branches where each branch focuses on learning features with different scales. Our proposed trident spatial attention can be formulated as:

$$\begin{aligned} TSA(f) = \sigma(Concat(&Conv_1(f') \\ &Conv_3(f') \\ &Conv_5(f')). \end{aligned} \tag{4}$$

where $Conv_1, Conv_3, Conv_5$ stand for convolutions with different dilation values.

To attentively aggregate multi-modal features, we follow the pipeline of the Cross-Reference Module (CRM) as suggested in DCF [12]. Formally, let R and D the paired RGB-D input for the fusion module, we first compute the modality-specific channel CA_r and CA_d, as well as the shared channel attention CA_f as follow:

$$\begin{aligned} &CA_r = CA(R); \quad CA_d = CA(D); \\ &CA_f = norm(max(CA_r, CA_d)); \end{aligned} \tag{5}$$

The vanilla CRM benefits from channel attention to realize the self- and cross-calibration before the feature fusion. We have:

$$\begin{aligned} CRM(R, D) = Concat(&CA_f \otimes CA_r \otimes R; \\ &CA_f \otimes CA_d \otimes D); \end{aligned} \tag{6}$$

We refer readers to the original paper [12] for more details on the cross-modal interaction. In our application, we replace the final concatenation with adaptive addition with respect to depth quality and form our $qCRM^{CA}$ as follows:

$$\begin{aligned} qCRM^{CA}(R, D) = &CA_f \otimes CA_r \otimes R + \\ &\lambda \cdot CA_f \otimes CA_d \otimes D; \end{aligned} \tag{7}$$

Moreover, we additionally design another branch where the CA is replaced by our proposed TSA. This new branch is termed as $qCRM^{TSA}$. We further learn two scalar values α and β to adaptively weight CRM^{TSA} with the original branch CRM with channel attention. Our adaptive fusion (AF) can be formulated as:

$$AF(R, D) = \alpha \cdot qCRM^{CA}(R, D) + \beta \cdot qCRM^{TSA}(R, D) \tag{8}$$

Finally, we merge the previous level output f_{i-1}, if any, and the current AF output with concatenation-convolution. To deal with the resolution, we apply max-pooling on f_{i-1} to preserve the most informative hierarchical features.

3.3. Architecture

In this paper, we propose a novel fusion design that can be easily adapted to any existing architecture. To compete with the state-of-the-art performance, we choose Res2Net [10] as our backbone to extract features. Our decoder is the same as SPNet [62]. Specifically, it consists of five-level RFB blocks [24]. Each block is skipped and connected with the fused encoded features. However, different from SPNet with a triple decoder to explicitly both modality-specific and shared features, we only maintain one decoder to decode our efficiently fused features. Our network is supervised by conventional IoU and BCE losses.

4. Experimental Validation

4.1. Datasets, Metrics and Training Settings

We follow previous works [6, 47, 12, 62] and train our model on the conventional training set which contains 1,485 samples from the NJU2K-train [15] and 700 samples from the NLPR-train [32]. For testing benchmarks, we observe that the depth quality within each dataset varies, which is mainly due to acquisition methods. Specifically, DES [3] contains 135 images of indoor scenes captured by a Kinect camera. SIP [5] provides a human dataset that contains 929 images captured by a mobile device. Therefore, these two datasets can be considered moderate with less noisy depths.

The remaining NLPR-test [32], NJU2K-test [15] and STERE [28] datasets are more challenging. NLPR-test [32] contains 300 natural images which are captured by a Kinect sensor. However, the images are obtained under different illumination conditions. NJU2K-test [15] contains 500 stereo image pairs from different sources such as the Internet and 3D movies. A number of depth maps are estimated through the optical flow method [38]. STERE [28] contains 1,000 stereoscopic images where the depths are estimated with

Table 1. Quantitative comparison with different fusion designs. We replace our fusion module with five SOTA fusion modules and retrain the new networks with the same training setting. ↑ (↓) denotes that the higher (lower) is better.

Dataset	Size↓	DES				NLPR				NJU2K				STERE				SIP			
Metric	(Δ Mb)	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$
BBS [6]	460(+95)	.015	**.946**	**.941**	.976	.023	.920	.923	.953	.035	.924	.915	.944	.040	.913	.902	.935	.053	.904	.877	.912
DFM[56]	495(+130)	.015	**.946**	**.941**	.974	.022	.919	.926	.956	.034	.922	.917	.943	.041	.909	.902	.933	.049	.909	.885	.919
CDI[52]	520(+155)	.023	.919	.914	.950	.024	.918	.921	.952	.035	.927	.915	.944	.036	.918	.910	.941	.055	.900	.870	.911
DCF[12]	336(**-29**)	.015	.944	.938	.976	**.020**	.927	**.931**	.960	.030	.930	.924	.949	.038	.913	.904	.937	.044	.913	.891	.928
SPNet[62]	593(+228)	.016	.944	.936	.973	.022	.924	.925	.956	.032	.928	.919	.945	.038	.913	.904	.938	.048	.907	.884	.921
MobSal[45]	723(+358)	**.015**	.945	.940	.976	.024	.924	.923	.955	.033	.926	.915	.945	.038	.913	.902	.937	**.042**	.915	.892	.930
Ours	365	**.015**	**.946**	**.941**	**.977**	**.020**	**.932**	**.931**	**.962**	**.029**	**.936**	**.926**	**.951**	**.035**	**.921**	**.911**	**.944**	**.042**	**.916**	**.893**	**.931**

SIFT flow method [20]. Due to the measurement or estimation error, these datasets contain more noisy depths. Therefore, to purely analyze the performance under different conditions, we additionally report the average metric (AvgMetric) for datasets with good quality depths and for datasets with more challenging depths.

To quantify the performance of our methods, we use conventional saliency metrics such as Mean Absolute Error, F-measure, S-measure, and E-measure. More details can be found in the supplementary material.

Our method is based on the Pytorch framework and is learned with a V100 GPU. The encoder is initialized with the pre-trained weights. For the 1-channel depth input, we replace the first convolution of backbone to feet with the depth size. The learning rate is initialized to $1e{-}4$ which is further divided by 10 every 60 epochs. We fix and resize the input RGB-D resolution to 352×352. During training, we adopt random flipping, rotating, and border clipping for data augmentation. The total training time takes around 5 hours with batch size 10 and epoch 100.

4.2. Comparison with SOTA fusion alternatives

We observe that existing works adopt different architectures, i.e., choice of backbones, design of decoder, supervision, training settings, etc. For example, light models [56, 45] always choose MobileNet [36] to extract features. Several works [30, 60, 34, 14] are based on VGG encoders [37] , while another group of models [56, 7] takes ResNet [11] as encoders. Recent works [62, 26] are based on more powerful backbones such as Res2Net [10] and ViT [4]. The choice of backbone will undoubtedly impact the final performance. Furthermore, the design of the decoder varies from one work to another. Several works are based on DenseASPP [51], while others are based on RFB [24]. Under the consideration of a fair comparison, we re-implement six SOTA fusion works under the same architecture. Specifically, we choose the same backbone, same decoder, loss, and same training settings as ours. The only difference between one model to another is in the fusion module. We refer readers to previous sections for more experimental details. Note that several fusion designs [45, 12] were initially applied only to certain layers. To fairly and purely analyze the fusion performance, we implement all the fusion modules at each scale as ours.

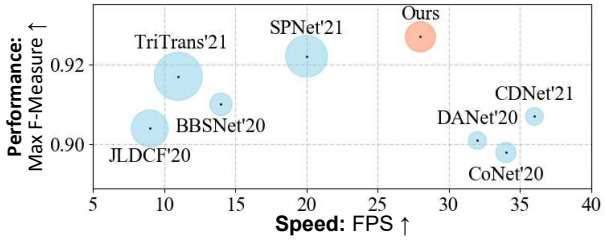

Figure 5. Average **Performance, Speed, and Model Size** of different methods on challenging datasets (NLPR, NJUK, STERE). The circle size denotes the model size. Note that better models are shown in the upper right corner (i.e., with a larger F-measure and larger FPS). Our method finds the best trade-off of the three measures. Methods with higher speed perform inferior, making our method both efficient and accurate.

Table 1 illustrates the quantitative comparison. We also report the model size of each embedded fusion module. $\Delta Size$ stands for the difference in model size compared to ours. It can be seen that our fusion strategy yields significantly better results compared to our counterparts. Compared to the lightest DCF fusion which only applies channel attention during feature fusion, we add additional spatial attention, yielding a slightly heavier model size (+29 Mb) but favorably improving the performance. Elsewise, our model size is significantly lighter compared to other counterparts, validating the effectiveness of our proposed fusion module.

4.3. Quantitative Comparison

Table 2 illustrates the quantitative comparison. For challenging datasets (NLPR, NJU2K, and STERE), our method performs favorably over the existing methods and sets a new state-of-the-art (SOTA) record, validating the superior robustness of our approach against depth bias. We further illustrate in Figure 5 the trade-off between model efficiency and SOTA performances. Compared to the current SOTA TriTrans [26] with 11 FPS and SPNet [62] with 20 FPS, our network achieves superior performance with higher inference speed, i.e., 28 FPS. For other datasets with less depth noise (DES and SIP), we also achieve competitive performance with almost halved the model size compared to the

Table 2. Quantitative comparison with state-of-the-art models. ↑ (↓) denotes that the higher (lower) is better. The best and second best are highlighted in **bold** and underline, respectively. We further report the average metric (AvgMetric) for datasets with more challenging depths and with less noisy depths.

Dataset	Size↓	Benchmarks with challenging depth																				Benchmarks with less noisy depth															
		NLPR					NJU2K					STERE					**AvgMetric**					DES					SIP					**AvgMetric**				
Metric	(Mb)	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$		$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$		$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$		$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$		$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$		$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$		$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	
$CPFP_{19}$ [59]	278	.036	.867	.888	.932	.053	.877	.878	.923	.051		.874	.879	.925	.049		.873	.880	.925		.038	.846	.872	.923	.064	.851	.850	.903	.060	.850	.852	.905				
$DMRA_{19}$ [33]	238	.031	.879	.899	.947	.051	.886	.886	.927	.047		.886	.886	.938	.045		.884	.888	.936		.030	.888	.900	.943	.085	.821	.806	.875	.078	.829	.817	.883				
$A2dele_{20}$ [34]	**116**	.029	.882	.898	.944	.051	.874	.871	.916	.044		.879	.878	.928	.043		.878	.879	.927		.029	.872	.886	.920	.070	.833	.828	.889	.060	.850	.852	.905				
$JLDCF_{20}$ [9]	548	.022	.916	.925	.962	.043	.903	.903	.944	.042		.901	.905	.946	.038		.904	.907	.948		.022	.919	.929	.968	.051	.885	.879	.923	.047	.889	.885	.928				
$CMMS_{20}$ [17]	546	.027	.896	.915	.949	.044	.897	.900	.936	.043		.893	.895	.939	.040		.894	.899	.939		.018	.930	.937	.976	.058	.877	.872	.911	.052	.883	.880	.918				
$CoNet_{20}$ [13]	162	.031	.887	.908	.945	.046	.893	.895	.937	.040		.905	.908	**.949**	.040		.898	.904	.945		.028	.896	.909	.945	.063	.867	.858	.913	.058	.870	.864	.917				
$DANet_{20}$ [60]	128	.028	.916	.915	.953	.045	.910	.899	.935	.043		.892	.901	.937	.041		.901	.902	.939		.023	.928	.924	.968	.054	.892	.875	.918	.050	.896	.881	.924				
$DASNet_{20}$ [57]	-	.021	.929	.929	-	.042	.911	.902	-	.037		.915	.910	-	.035		.916	.910	-		.023	.928	.908	-	-	-	-	-	-	-	-	-				
$HDFNet_{20}$ [30]	308	.031	.839	.898	.942	.051	.847	.885	.920	.039		.863	.906	.937	.041		.854	.898	.933		.030	.843	.899	.944	.050	.904	.878	.920	.047	.896	.880	.923				
$BBSNet_{20}$ [6]	200	.023	.918	.930	.961	.035	.920	.921	.949	.041		.903	.908	.942	.036		.910	.915	.947		.021	.927	.933	.966	-	-	-	-	-	-	-	-				
DCF_{21} [12]	435	.021	.891	-	.957	.035	.902	-	.924	.039		.885	-	.927	.034		.890	-	.931		-	-	-	-	.051	.875	-	.920	-	-	-	-				
$D3Net_{21}$ [5]	518	.030	.897	.912	.953	.041	.900	.900	.950	.046		.891	.899	.938	.041		.894	.901	.943		.031	.885	.898	.946	.063	.861	.860	.909	.058	.864	.864	.913				
$DSA2F_{21}$ [39]	-	.024	.897	.918	.950	.039	.901	.903	.923	.036		.898	.904	.933	.034		.898	.906	.933		.021	.896	.920	.962	-	-	-	-	-	-	-	-				
$TriTrans_{21}$ [26]	927	.020	.923	.928	.960	.030	.926	.920	.925	**.033**		.911	.908	.927	.030		.917	.914	.931		.014	.940	.943	.981	.043	.898	.886	.924	.039	-	.893	.931				
$CDINet_{21}$ [52]	217	.024	.916	.927	-	.035	.922	.919	-	.041		.903	.906	-	.036		.910	.913	-		-	-	-	-	-	-	-	-	-	-	-	-				
$SPNet_{21}$[62]	702	.021	.925	.927	.959	**.028**	.935	.925	**.954**	.037		.915	.907	.944	.031		.922	.915	**.949**		**.014**	**.950**	**.945**	**.980**	.043	**.916**	**.894**	.930	.039	**.920**	**.900**	**.936**				
RFNet (ours)	364	**.020**	**.932**	**.931**	**.962**	.029	**.936**	**.926**	.951	.035		**.921**	**.911**	.944	**.030**		**.927**	**.918**	.948		.015	.946	.941	.977	**.042**	**.916**	.893	**.931**	**.038**	.919	.899	**.936**				

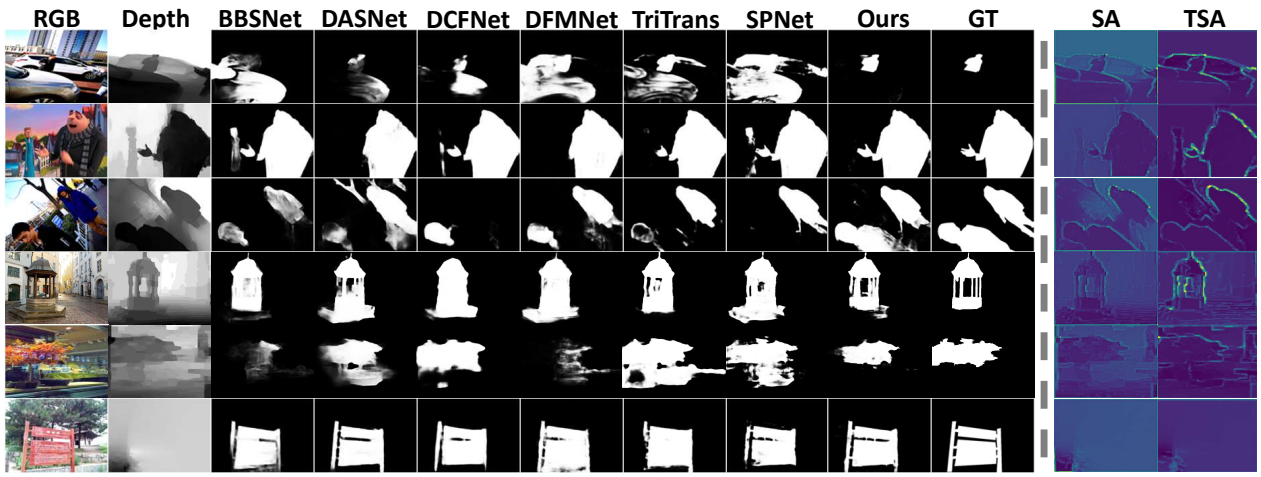

Figure 6. **Qualitative comparison**. We also illustrate the depth features enhanced by vanilla SA and by our proposed TSA, respectively. Our work yields more boundary activation compared to the counterpart. Better to zoom in.

current SOTA SPNet [62]. Note that both SPNet and ours adopt Res2Net50 [10] as the backbone. Thus, our performance can be contributed to our proposed fusion solely.

4.4. Qualitative Comparison

Figure 6 presents the generated saliency maps of different methods on challenging cases such as single or multiple humans, clustered foreground-background, and low-quality depth. It can be seen that our methods consistently reason about saliency masks closer to the ground truth. We further illustrate the comparison between depth feature maps enhanced with our proposed spatial attention (TSA) and with the counterpart (TA). We can visualize that our attention is more sensitive to camera distances and can better segment object regions. This can be contributed to our trident branches with different scales. Furthermore, our attention yields more activation on the boundary, facilitating the network to better leverage geometric for saliency detection.

Finally, we illustrate in Figure 7 the histogram for our layer-wise attention. We particularly choose λ_1 and λ_5 to facilitate the understanding of the trade-off between early and late fusion. We can observe that while depths are of low quality, our LWA assigns more weights for late fusion (with low λ_1 value and high λ_5 value). While depths are of good quality, our LWA assigns more weights for early fusion (with high λ_1 value and low λ_5 value). This observation is consistent with previous studies [52, 30, 12, 7] with discrepant fusion. We hope our analysis of layer-wise attention can inspire future adaptive fusion works.

4.5. Distribution of Spatial and Channel Attention

Since we propose an adaptive weighting strategy to merge our spatial attention (TSA) and channel attention (CA), we illustrate in Figure 8 the distribution of weights of each attention at different stages of the network. We can observe that TSA and CA contribute differently with respect to the network depth. At layer 1 (L_1), the network assigns more weights to TSA, which can be explained by the sig-

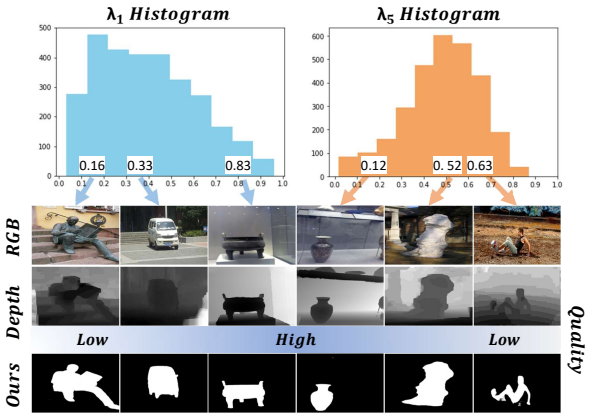

Figure 7. **Trade-off between early and late fusion**. Our layer-wise attention can adaptively model the depth contribution during feature fusion. While are of low quality, we assign less weight to early fusion since the noisy geometric cues are difficult to be exploited. Meanwhile, we assign more weight to late fusion to leverage the multi-modal semantic cues for feature fusion.

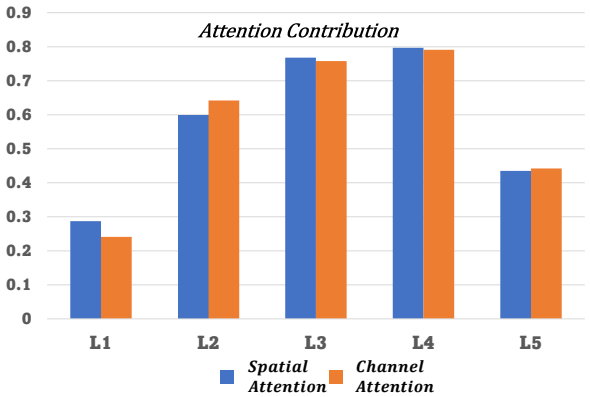

Figure 8. **Attention contribution**. $L_1, ...L_5$ stands for the different layers. We realize attention fusion through $\alpha \cdot CA + \beta \cdot TSA$.

nificant spatial resolution of the features. For deeper layers, TSA and CA tend to play a similar role at each stage to enhance the feature modeling with equal importance from both spatial and channel cues. However, we can observe that attention from different layers contribute differently to the final output, i.e., attention from the third layer (L_3) and the fourth layer (L_4) contribute more compared to the first two layers ($L_1 - L_2$) and the last layer (L_5). The difference with respect to the network depth is also consistent with previous work [29] and to our layer-wise attention that shallow and deep layers play different roles for feature fusion.

4.6. Ablation Study

In this section, we conduct an ablation study to validate the effectiveness of each proposed component. The quantitative result of each combination can be found in Table 3. To analyze the effectiveness of our trident spatial attention (TSA), we replace ours with vanilla spatial attention [44] and observe a dropped performance. This is mainly due to the limited receptive field of vanilla attention that assumes a local correlation between different features. In contrast, our TSA can significantly improve performance by leveraging contextualized awareness. The boosted performance on the aforementioned datasets validates the design of our TSA.

Table 3. Ablation study on key components. B stands for the baseline performance where RGB-D features are merged through simple addition without any form of attention.

B	CRM				DFM	LWA	Size	Overall Metric			
	CA	TSA	SA	α, β	([56])		Mb↓	$M \downarrow$	$F \uparrow$	$S \uparrow$	$E \uparrow$
✓							**305**	.039	.915	.904	.935
✓	✓						336	.035	.918	.907	.940
✓	✓	✓					364	.035	.923	.910	.943
✓	✓		✓				363	.035	.920	.908	.941
✓	✓	✓		✓			364	.034	**.924**	.910	.943
✓	✓	✓		✓	✓		364	.035	.921	.908	.941
✓	✓	✓		✓		✓	365	**.033**	**.924**	**.911**	**.944**

We also conduct experiments by replacing our LWA with the concurrent DFM presented in DFMNet [56]. We can observe that the performance significantly degrades. The difference between DFM and ours is in the manner to compute the similarity matrix. Specifically, DFM assumes a perfect alignment between multi-modalities and realizes a pixel-wise matrix multiplication, while we leverage the non-local attention with flattened vectors to compute the similarity.

5. Conclusion

In this paper, we proposed a novel fusion architecture for RGB-D saliency detection. Different from previous works, we improve the robustness against inaccurate and misaligned depth inputs. Specifically, we proposed a novel layer-wise attention to explicitly leverage the depth quality by learning the best trade-off between early and late fusion. Furthermore, we improved the vanilla spatial attention to a broader context, yielding a simple yet efficient mechanism to address the depth misalignment problem. Extensive comparisons on benchmark datasets validate the effectiveness and robustness of our approach compared to the state-of-the-art alternatives. Our method also sets new records on challenging datasets with smaller model sizes. The method developed in this paper can potentially be used for other tasks, such as semantic segmentation and object detection, in a similar setting of RGB-D inputs in a robust manner.

Acknowledgements

We gratefully acknowledge Zhuyun Zhou and Deng-Ping Fan for discussion and proofreading. 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 and the French "Investissements d'Avenir" program ISITE-BFC (ANR-15-IDEX-0003).

References

[1] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media (CVM)*, 5(2):117–150, 2019. 2

[2] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[3] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS)*, 2014. 5

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representation (ICLR)*, 2021. 6

[5] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems (TNNLS)*, 32(5):2075–2089, 2021. 1, 3, 5, 7

[6] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling 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. 1, 2, 3, 4, 5, 6, 7

[7] Xian Fang, Jinshao Zhu, Xiuli Shao, and Hongpeng Wang. GroupTransNet: Group transformer network for RGB-D salient object detection. *arXiv preprint arXiv:2203.10785*, 2022. 2, 3, 4, 6, 7

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun 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. 2, 7

[10] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(2):652–662, 2021. 5, 6, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[12] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 4, 5, 6, 7

[13] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 7

[14] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:3376–3390, 2021. 6

[15] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE International Conference on Image Processing (ICIP)*, 2014. 5

[16] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 2

[17] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[18] Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:3528–3542, 2021. 3

[19] Guibiao Liao, Wei Gao, Qiuping Jiang, Ronggang Wang, and Ge Li. Mmnet: Multi-stage and multi-scale fusion network for rgb-d salient object detection. In *Proceedings of the 28th ACM international conference on multimedia (ACMMM)*, 2020. 3

[20] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):978–994, 2011. 6

[21] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin 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. 1

[22] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4

[23] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[24] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5, 6

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Pro-*

ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 3

[26] Zhengyi Liu, Wang Yuan, Zhengzheng Tu, Yun Xiao, and Bin 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. 1, 2, 3, 4, 6, 7

[27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, 2016. 4

[28] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[29] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution. *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 8

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

[31] Danda Pani Paudel, Cédric Demonceaux, Adlane Habed, and Pascal Vasseur. 2D-3D synchronous/asynchronous camera fusion for visual odometry. *Autonomous Robots*, 43(1):21–35, 2019. 1

[32] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 5

[33] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7

[34] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan 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. 1, 2, 6, 7

[35] Wang Qilong, Wu Banggu, Zhu Pengfei, Li Peihua, Zuo Wangmeng, and Hu Qinghua. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *The IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 6

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, (ICLR)*, 2015. 6

[38] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 5

[39] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi 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. 2, 7

[40] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2020. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4

[42] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2021. 2

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 4

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 3, 5, 8

[45] Yu-Huan Wu, Yun Liu, Jun Xu, Jia-Wang Bian, Yu-Chao Gu, and Ming-Ming Cheng. Mobilesal: Extremely efficient rgb-d salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 6

[46] Zongwei Wu, Guillaume Allibert, Christophe Stolz, and Cédric Demonceaux. Depth-Adapted CNN for RGB-D cameras. In *Asian conference on computer vision (ACCV)*, 2020. 2

[47] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Modality-guided subnetwork for salient object detection. In *2021 International Conference on 3D Vision (3DV)*, 2021. 1, 2, 5

[48] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Depth-adapted CNNs for RGB-D semantic segmentation. *arXiv preprint arXiv:2206.03939*, 2022. 2

[49] Zhe Wu, Li Su, and Qingming 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. 1

[50] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[51] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[52] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam 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. 1, 2, 3, 4, 5, 6, 7

[53] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. RGB-D saliency detection via cascaded mutual information minimization. In *Poceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[54] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan 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. 2

[55] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[56] Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun 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. 1, 2, 3, 4, 6, 8

[57] 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. 1, 2, 3, 4, 7

[58] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le 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. 2

[59] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le 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. 7

[60] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei 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. 2, 6, 7

[61] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. *Computational Visual Media (CVM)*, pages 1–33, 2021. 2

[62] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving RGB-D saliency detection. In *Poceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 6, 7