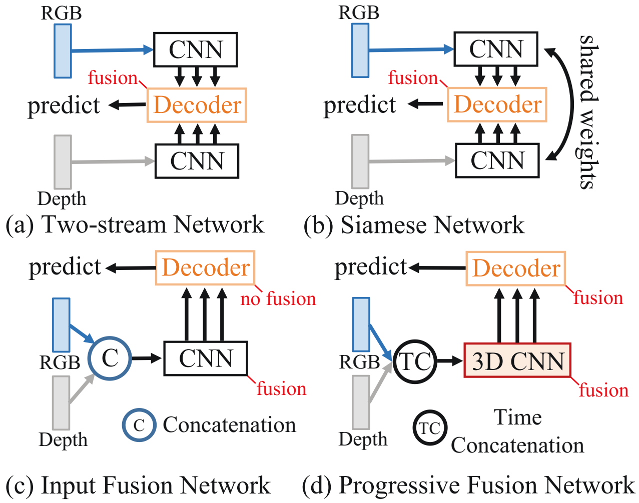
RGB-D Salient Object Detection via 3D Convolutional Neural Networks

Qian Chen[1], Ze Liu[1], Yi Zhang[2], Keren Fu[3,4]*, Qijun Zhao[3,4], Hongwei Du[1]
[1]School of Information Science and Technology, University of Science and Technology of China
[2]Institut National des Sciences Appliquées de Rennes
[3]College of Computer Science, Sichuan University
[4]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University
{poly,liuze}@mail.ustc.edu.cn, yi.zhang1@insa-rennes.fr, {fkrsuper,qjzhao}@scu.edu.cn, duhw@ustc.edu.cn

Abstract

RGB-D salient object detection (SOD) recently has attracted increasing research interest and many deep learning methods based on encoder-decoder architectures have emerged. However, most existing RGB-D SOD models conduct feature fusion either in the single encoder or the decoder stage, which hardly guarantees sufficient cross-modal fusion ability. In this paper, we make the first attempt in addressing RGB-D SOD through 3D convolutional neural networks. The proposed model, named *RD3D*, aims at pre-fusion in the encoder stage and in-depth fusion in the decoder stage to effectively promote the full integration of RGB and depth streams. Specifically, *RD3D* first conducts pre-fusion across RGB and depth modalities through an inflated 3D encoder, and later provides in-depth feature fusion by designing a 3D decoder equipped with rich back-projection paths (RBPP) for leveraging the extensive aggregation ability of 3D convolutions. With such a progressive fusion strategy involving both the encoder and decoder, effective and thorough interaction between the two modalities can be exploited and boost the detection accuracy. Extensive experiments on six widely used benchmark datasets demonstrate that *RD3D* performs favorably against 14 state-of-the-art RGB-D SOD approaches in terms of four key evaluation metrics. Our code will be made publicly available: https://github.com/PPOLYpubki/RD3D.



(a) Two-stream Network (b) Siamese Network

(c) Input Fusion Network (d) Progressive Fusion Network

Figure 1: Categorization of existing models. Note that (a)-(c) conduct feature fusion either in the encoder or the decoder stage, while our model (d) adopts progressive fusion involving both the encoder and decoder stages.

Introduction

Salient object detection (SOD) aims to imitate the human visual system on detecting objects or areas that attract human attention (Jiang et al. 2020; Fan et al. 2018a; Zhao et al. 2019b). SOD has a wide range of applications in many tasks, such as object segmentation and recognition (Han et al. 2005; Li, Zhou, and Yang 2011), video detection (Li et al. 2019; Fan et al. 2019), content-related image and video compression (Itti 2004; Guo and Zhang 2009) as well as tracking (Zhang et al. 2020d). Although SOD has been advanced notably by deep learning techniques (Wang et al. 2019), single-modal SOD still faces many problems, such as weak appearance differences in the foreground and background regions, complex foreground and background, *etc*.

In recent years, an increasing number of RGB-D SOD models have emerged to address these challenges of single-modal SOD for more accurate detection performance (Zhang et al. 2021). Although encouraging results have been obtained, we notice that existing models conduct feature fusion either in the single encoder or the decoder stage, which may hardly guarantee sufficient cross-modal fusion ability. As shown in Fig. 1, these models can be divided into three categories according to how they extract and fuse cross-modal features. In the first category (Fig. 1 (a)), the models (Fan et al. 2020b; Pang et al. 2020; Liu, Zhang, and Han 2020; Piao et al. 2020; Zhang et al. 2020c) extract features from RGB and depth maps independently, and conduct feature maps fusion of the two modalities in the decoder. To achieve effective cross-modal fusion, the authors tend to elaborately design complex or special modules for simultaneous fusion and decoding. The second category (Fu et al. 2020a; Li, Liu, and Ling 2020) (Fig. 1 (b)) uses a Siamese

*Corresponding author: Keren Fu (*fkrsuper@scu.edu.cn*) network as an encoder to extract features from RGB and depth. Although the encoder network is shared across different modalities, however, it is still dedicated to feature extraction similar to Fig. 1 (a) and no fusion behavior is conducted in the encoder. The third category of models (Zhao et al. 2020; Fan et al. 2020a; Song et al. 2017; Liu et al. 2019) (Fig. 1 (c)) adopt the "input fusion" strategy, which concatenates RGB and depth across channel dimension before feeding them to the encoder. In this case, the main role of fusion is played by the encoder since all the ingredients fed to the decoder are already-fused features, making the decoder infeasible to conduct explicit cross-modal fusion.

Considering that feature extraction and fusion is crucial in such an encoder-decoder architecture for the RGB-D SOD task, the aforementioned models have not fully investigated the feature aggregation potentials in both the encoder and decoder. Inspired by the success of 3D convolutional neural networks (CNNs) in aggregating extensive feature information for space-time processing (*e.g.*, video recognition (Feichtenhofer 2020), action localization (Gu et al. 2018)) where 3D CNNs often serve as encoders, we propose to treat the depth modality as another "time state" of the RGB one and aggregate information of the two modalities through 3D CNNs. To the best of knowledge, our work is *the first attempt that addresses RGB-D SOD through 3D CNNs*, attributed to which RGB and depth information can be mutually enhanced meanwhile making explicit fusion in the decoder possible. Another advantage is that due to the inner fusion behavior of 3D convolutions, dedicated or sophisticated modules for cross-modal fusion are *no longer required*. The proposed novel model, named *RD3D* (short for **R**GB-**D** **3D** CNN detector for SOD), first conducts pre-fusion across RGB and depth modalities through an inflated 3D encoder. Then, the obtained pre-fused RGB and depth features are fed to a 3D decoder for further in-depth fusion. The 3D decoder incorporates rich back-projection paths (RBPP) in order to better leverage the extensive aggregation ability of 3D convolutions. Therefore, both the encoder and decoder of *RD3D* are 3D CNNs-based and they both involve cross-modal fusion in a progressive manner (Fig. 1 (d)). Our work has three main contributions:

- We exploit the idea of pre-fusion in the encoder stage and show how it is beneficial to the final performance. We propose to tackle this by 3D CNNs, which can fuse the cross-modal features effectively without requiring dedicated or sophisticated modules.
- We design a 3D decoder that incorporates rich back-projection paths (RBPP) in order to better leverage the extensive aggregation ability of 3D convolutions. Such a 3D decoder makes the proposed *RD3D* a fully 3D CNNs-based model and also the first 3D CNNs-based model for the RGB-D SOD task.
- We show that *RD3D*, which is the first 3D CNNs-based model for RGB-D SOD, surpasses 14 state-of-the-art (SOTA) methods by a notable margin on the six widely used benchmark datasets.

Related Work

Deep-based RGB-D Models. Existing deep models can be divided into three classes according to the stage of fusion: early-fusion (Peng et al. 2014; Song et al. 2017), middle-fusion (Feng et al. 2016; Fu et al. 2020a,b; Zhang et al. 2020b; Piao et al. 2019) and late-fusion (Fan, Liu, and Sun 2014). By contrast, as shown in Fig. 1, this paper elaborately divides current methods into four categories according to how they extract and fuse cross-modal features. Among the first category named the two-stream network (Fig. 1 (a)), (Han et al. 2017) utilized a CNN network to extract information from the two modalities in the backbone stage, and then fused such deep representations from multi-views via a fully connected layer. (Piao et al. 2019) proposed a novel depth-induced multi-scale recurrent attention network, which extracted features respectively from RGB and depth maps and then input them to depth refinement blocks for integration. (Chen et al. 2020) utilized separate CNNs to extract features from RGB and depth modalities. The resulting hint map is then utilized to enhance the depth map, which suppresses the noise and sharpens the object boundary. The second category is the Siamese network (Fig. 1 (b)). Fu *et al.* (Fu et al. 2020a) and Li *et al.* (Li, Liu, and Ling 2020) first adopted a Siamese network with shared weights for the RGB/depth stream during independent feature extraction. The third category is called the input fusion network (Fig. 1 (c)). (Huang, Shen, and Hsiao 2018) and (Liu et al. 2019) concatenated RGB and depth maps to formulate a four-channel input, which was fed to a single-stream CNN. DANet proposed by Zhao *et al.* (Zhao et al. 2020) fused bi-modal information in the input stage, and meanwhile depth maps played a guidance role in the decoder stage.

In general, the above representative works do their utmost to explore: 1) effective utilization of depth information, and 2) comprehensive fusion of RGB and depth cues. Unfortunately, they have the limitation that feature aggregation potentials in both the encoder and decoder are not fully leveraged. Complete survey of models in this field can be found in (Zhou et al. 2021). Different from existing models, we propose the 3D CNNs-based progressive fusion scheme (Fig. 1 (d)) towards a new perspective of multi-modal feature extraction and fusion.

3D CNNs. 3D CNNs are influential in many fields, such as video processing (Ji et al. 2012; Tran et al. 2015), medical image processing (Balakrishnan et al. 2019) and point cloud processing (Zhou and Tuzel 2018). Balakrishnan *et al.* (Balakrishnan et al. 2019) applied 3D convolutions to extract features from image volumes in the encoder stage and utilized a 3D CNN-based decoder to transform features on finer spatial scales, enabling precise anatomical alignment. By using 3D convolutions, Zhou *et al.* (Zhou and Tuzel 2018) extracted features from 3D voxels for point cloud-based 3D object detection. The above methods use 3D convolutions to handle data resided in the 3D space, but 3D convolutions can also process data in multi-domain. Ji *et al.* (Ji et al. 2012) applied them to extract features in spatial and temporal domains from video data to capture motion information. To the best of our knowledge, we are the first to investigate 3D convolutions for RGB-D saliency detection.

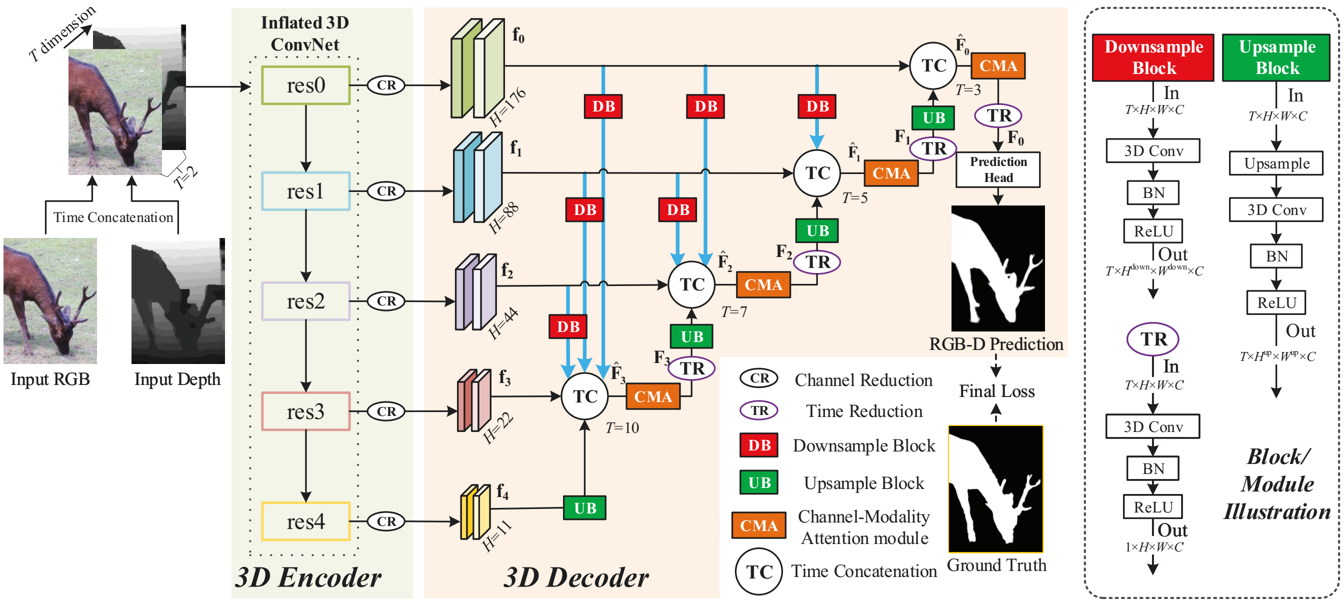


Figure 2: Block diagram of the proposed *RD3D* scheme for RGB-D SOD. H denotes the spatial resolution of output feature maps at each level, and T denotes the temporal dimension. Definitions of $\mathbf{f}_i$, $\hat{\mathbf{F}}_i$, and $\mathbf{F}_i$ can be found in Eq. (2) and Eq. (3).

Methodology

Big Picture

The overall architecture of the proposed *RD3D* is shown in Fig. 2. It follows the typical encoder-decoder architecture and is composed of a 3D encoder and a 3D decoder. The 3D encoder is basically a ResNet/VGG-like backbone which is extended by 3D convolutions. It aims at cross-modal feature pre-fusion while its outputs are modality-aware multi-level features. On the other hand, the 3D decoder decodes features by 3D convolutions. It follows the typical UNet-like top-down fashion but incorporates rich back-projection paths (RBPP, the blue line arrows in Fig. 2) as well as channel-modality attention modules (CMA, the orange modules in Fig. 2). After the final decoding by 3D convolutions, the decoder outputs a prediction map highlighting salient object(s). Noting that attributed to the extensive aggregation ability of 3D convolutions, no any explicit cross-modal fusion modules are used in Fig. 2.

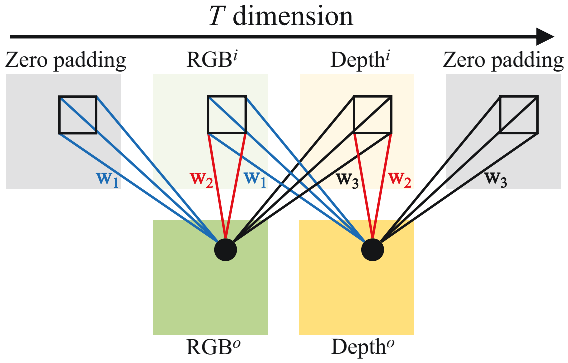


Figure 3: Visualization of 3D convolution in the "T" dimension, where the corresponding kernel size is 3. The superscript "i/o" means input/output features of 3D convolution.

3D Encoder

As shown in Fig. 2, given an RGB image and a single-channel depth map, we first normalize the depth map into intervals $[0, 255]$ and then replicate it into three channels. Hereafter, we follow (Wang et al. 2018) and denote the dimension of a tensor as $T \times H \times W \times C$, where "$T$" refers to the temporal dimension and "H", "W", "C" mean the height, width, and channels, respectively. We stack the RGB image ($H \times W \times C$) and the corresponding depth map ($H \times W \times C$) to form a 4D tensor as the input of our 3D encoder, where $T = 2$ and $C = 3$. We adopt an inflated 3D ResNet (Carreira and Zisserman 2017) as our encoder, which replaces all 2D convolutions in the conventional ResNet (He et al. 2016) with 3D convolutions, and the kernel sizes for the "T" dimension are set as 3 for all the 3×3 3D convolutions, with padding, stride, output dimension being 1, 1 and 2, respectively. Computation in the "T" dimension of a 3D convolutional layer thus can be visualized in Fig. 3 and is equivalent to the formulations below:

$$\begin{aligned} \mathbf{R}^o &= \mathbf{w_2} * \mathbf{R}^i + \mathbf{w_3} * \mathbf{D}^i, \\ \mathbf{D}^o &= \mathbf{w_1} * \mathbf{R}^i + \mathbf{w_2} * \mathbf{D}^i, \end{aligned} \tag{1}$$

where $\mathbf{w_1}$, $\mathbf{w_2}$ and $\mathbf{w_3}$ represent the three temporal weight slices of the 3D kernel. $\mathbf{R}^i$ and $\mathbf{D}^i$ denote the input RGB and depth feature slices, respectively, whereas $\mathbf{R}^o$ and $\mathbf{D}^o$ denote the output ones. "$*$" means the 2D convolution operation. One can see that the inner fusion property of 3D convolutions helps fuse the RGB and depth information, where RGB and depth cues are mutually enhanced by each other

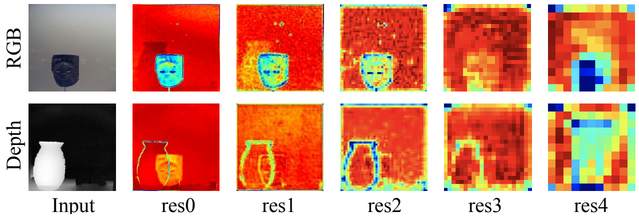


Figure 4: Modality-aware hierarchical features in each temporal slice. To make the impact of pre-fusion more visually obvious, we intentionally feed to our encoder RGB and depth images that are not matched. As a result, explicit fusion behavior can be observed.

when passing through a 3D convolutional layer. So, progressive fusion is achieved by using successive 3D convolutions.

Also note that the output number in the "T" dimension of our encoder is fixed as 2 under the particular consideration that there are only two modalities in our problem, namely RGB and depth. Although there exist other temporal designs of 3D kernels, Eq. (1) is adequate to reflect our idea of using 3D CNNs. Specifically, in Eq. (1) RGB and depth cues are preserved by shared weights $\mathbf{w_2}$, and meanwhile each one is enhanced by the other by learnable weights $\mathbf{w_1}/\mathbf{w_3}$. This achieves certain modality-aware individuality as well as cross-modal fusion, leading to the term "*pre-fusion*". Fig. 4 visualizes either temporal slice of feature maps at different levels, from which it is observed that information between the two modalities are cleverly integrated, but they are not the same. Finally, as shown in Fig. 2, the yielded modality-aware multi-level features whose temporal dimensions equal to 2 are fed to channel reduction (CR) modules to reduce their channels to a fixed smaller number (while the other dimensions are unchanged), *i.e.* 32 in practice, for the subsequent decoding. This is to reduce computation load as well as memory usage.

Inspired by (Carreira and Zisserman 2017; Feichtenhofer, Pinz, and Wildes 2016; Girdhar et al. 2018), we propose to initialize our 3D encoder in a centralized strategy using ImageNet pre-trained weights of the 2D ResNet, namely using such 2D weights to initialize the central slice $\mathbf{w_2}$ of a 3D kernel while setting other slices to 0, *i.e.*, $\mathbf{w_1} = \mathbf{w_3} = 0$. This strategy is equivalent to using a shared 2D ResNet to process RGB and depth at the beginning, which exactly coincides with the recent idea of using Siamese network for RGB-D SOD (Fu et al. 2020a).

3D Decoder with Rich Back-Projection Paths

As shown in the decoder part of Fig. 2, the channel-reduced features at each spatial resolution is aggregated with those at other resolutions in a hierarchical way. Inspired by but different from the widely employed UNet-like top-down fashion, which only considers *upsampling* low-resolution features to incorporate with high-resolution ones for refinement, we propose to combine additional *downsampling* flows from high-resolution features to low-resolution ones, denoted by the blue line arrows in Fig. 2, to leverage the extensive aggregation ability of 3D convolutions. Such downsampling flows transport rich feature information from the higher resolutions to the lower resolutions, enriching high-level feature representation. Note that besides the classical UNet architecture, this also contrasts to the existing technique (Hou et al. 2017) whose short connections transport information from high-level to low-level, since we transport in the opposite direction. We call our this method Rich Back-Projection Paths (RBPP). Another important reason of using RBPP is that, 3D convolutions will be more memory- and computation-efficient when used in such a decoder than in RBPP's counterparts that transport features in the opposite direction, like in (Hou et al. 2017).

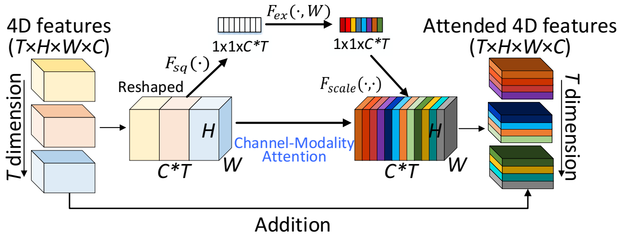


Figure 5: Proposed 3D channel-modality attention module that attends on both channel and temporal dimensions.

To be more specific in Fig. 2, for the ith level, we use a series of downsampling blocks to back-project features from all higher resolutions and meanwhile use an upsampling block to upsample the nearby aggregated feature outputs. The downsampling block is composed of a $1 \times 3 \times 3$ 3D convolutional layer, a BN layer, and a ReLU layer. In contrast, the upsampling block is composed of a bilinear upsampling layer and a $1 \times 3 \times 3$ 3D convolutional layer, followed by a BN layer and a ReLU layer. Note that both the downsampling and upsampling blocks will keep the temporal dimension number unchanged. Below, we denote the two blocks as $DB(\cdot)$ and $UB(\cdot)$, respectively. The feature computation at the ith level ($i \in \{0, 1, 2, 3\}$) is formulated as:

$$\hat{\mathbf{F}}_i = TConcat(\mathbf{f}_i, DB(\mathbf{f}_{i-1})...DB(\mathbf{f}_0), UB(\mathbf{F}_{i+1})) \quad (2)$$

$$\mathbf{F}_i = TR(CMA(\hat{\mathbf{F}}_i)), \quad (3)$$

where $TConcat(\cdot)$ means time concatenation (*i.e.*, concatenating in the temporal axis), $\mathbf{f}_i$ means the ith-level reduced feature tensor after the CR module in the encoder, $\mathbf{F}_{i+1}$ is the nearby feature outputs computed at the $(i+1)$th level, $TR(\cdot)$ denotes a temporal reduction operation which reduces the temporal dimension number to 1 as shown in Fig. 2, and $CMA(\cdot)$ denotes the Channel-Modality Attention module introduced below. $\hat{\mathbf{F}}_i$ denotes the intermediate features whereas $\mathbf{F}_i$ is the final feature outputs at the ith level. Note that we set $\mathbf{F}_4 = \mathbf{f_4}$, and the kernel sizes of the 3D convolutions in $TR(\cdot)$ vary from $10 \times 1 \times 1$ (when $i = 3$) to $3 \times 1 \times 1$ (when $i = 0$). After $\mathbf{F}_0$ is obtained, a prediction head consisting of a $(1 \times 1 \times 1, 1)$ convolutional layer and a Sigmoid layer is used to get the final prediction map.

Channel-Modality Attention Module. For generally enhancing features in a 3D decoder, we propose the Channel-Modality Attention module (CMA), which is inspired by

	Metric	AFNet	CTMF	PCF	MMCI	CPFP	D3Net	DMRA	SSF	A2dele	JLDCF	UCNet	CoNet	cmMS	DANet	**RD3D**
NJU2K	$S_\alpha \uparrow$	0.772	0.849	0.877	0.858	0.879	0.893	0.886	0.899	0.869	0.903	0.897	0.895	0.900	0.899	**0.916**
	$F_\beta^{\max} \uparrow$	0.775	0.845	0.872	0.852	0.877	0.887	0.886	0.896	0.873	0.903	0.895	0.893	0.897	0.898	**0.914**
	$E_\phi^{\max} \uparrow$	0.853	0.913	0.924	0.915	0.926	0.930	0.927	0.935	0.916	0.944	0.936	0.937	0.936	0.935	**0.947**
	$\mathcal{M} \downarrow$	0.100	0.085	0.059	0.079	0.053	0.051	0.051	0.043	0.051	0.043	0.043	0.047	0.044	0.045	**0.036**
NLPR	$S_\alpha \uparrow$	0.799	0.860	0.874	0.856	0.888	0.905	0.899	0.914	0.881	0.925	0.920	0.908	0.915	0.915	**0.930**
	$F_\beta^{\max} \uparrow$	0.771	0.825	0.841	0.815	0.867	0.885	0.879	0.896	0.881	0.916	0.903	0.887	0.896	0.903	**0.919**
	$E_\phi^{\max} \uparrow$	0.879	0.929	0.925	0.913	0.932	0.945	0.947	0.953	0.945	0.962	0.956	0.945	0.949	0.953	**0.965**
	$\mathcal{M} \downarrow$	0.058	0.056	0.044	0.059	0.036	0.033	0.031	0.026	0.028	0.022	0.025	0.031	0.027	0.028	**0.022**
STERE	$S_\alpha \uparrow$	0.825	0.848	0.875	0.873	0.879	0.889	0.886	0.893	0.879	0.905	0.903	0.908	0.895	0.901	**0.911**
	$F_\beta^{\max} \uparrow$	0.823	0.831	0.860	0.863	0.874	0.878	0.886	0.889	0.879	0.901	0.899	0.905	0.893	0.892	**0.906**
	$E_\phi^{\max} \uparrow$	0.887	0.912	0.925	0.927	0.925	0.929	0.938	0.936	0.928	0.946	0.944	**0.949**	0.939	0.937	0.947
	$\mathcal{M} \downarrow$	0.075	0.086	0.064	0.068	0.051	0.054	0.047	0.044	0.044	0.042	0.039	0.040	0.043	0.043	**0.037**
RGBD135	$S_\alpha \uparrow$	0.770	0.863	0.842	0.848	0.872	0.904	0.900	0.904	0.884	0.929	0.934	0.909	0.931	0.924	**0.935**
	$F_\beta^{\max} \uparrow$	0.728	0.844	0.804	0.822	0.846	0.885	0.888	0.884	0.870	0.919	**0.930**	0.895	0.922	0.914	0.929
	$E_\phi^{\max} \uparrow$	0.881	0.932	0.893	0.928	0.923	0.946	0.943	0.941	0.920	0.968	**0.976**	0.945	0.970	0.966	0.972
	$\mathcal{M} \downarrow$	0.068	0.055	0.049	0.065	0.038	0.030	0.030	0.026	0.029	0.022	0.019	0.028	0.019	0.023	**0.019**
DUTLF-D	$S_\alpha \uparrow$	0.468	0.831	0.801	0.791	0.749	0.775	0.889	0.915	0.885	0.913	0.863	0.919	0.912	0.899	**0.932**
	$F_\beta^{\max} \uparrow$	0.357	0.823	0.771	0.767	0.718	0.742	0.898	0.924	0.892	0.916	0.857	0.927	0.914	0.906	**0.939**
	$E_\phi^{\max} \uparrow$	0.638	0.899	0.856	0.859	0.811	0.834	0.933	0.951	0.930	0.949	0.904	0.956	0.943	0.940	**0.960**
	$\mathcal{M} \downarrow$	0.229	0.097	0.100	0.113	0.099	0.097	0.048	0.033	0.042	0.039	0.056	0.033	0.037	0.043	**0.031**
SIP	$S_\alpha \uparrow$	0.720	0.716	0.842	0.833	0.850	0.864	0.806	0.874	0.826	0.879	0.875	0.858	0.867	0.875	**0.885**
	$F_\beta^{\max} \uparrow$	0.712	0.694	0.838	0.818	0.851	0.861	0.821	0.880	0.832	0.885	0.879	0.867	0.871	0.876	**0.889**
	$E_\phi^{\max} \uparrow$	0.819	0.829	0.901	0.897	0.903	0.910	0.875	0.921	0.890	0.923	0.919	0.913	0.907	0.918	**0.924**
	$\mathcal{M} \downarrow$	0.118	0.139	0.071	0.086	0.064	0.063	0.085	0.053	0.070	0.051	0.051	0.063	0.060	0.054	**0.048**

Table 1: Quantitative SOD results in terms of S-measure (S_α), maximum F-measure ($F_\beta^{\max}$), maximum E-measure ($E_\phi^{\max}$) and mean absolute error ($\mathcal{M}$). Six widely used benchmark datasets are employed in the evaluation. ↑/↓ denotes that a larger/smaller value is better. The best results are highlighted in bold.

the squeeze-excitation attention block (Hu, Shen, and Sun 2018). The underlying purpose is to learn different attention weights *considering both channel and temporal dimensions*. As shown in Fig. 5, suppose the input is a 4D tensor with dimension $T \times H \times W \times C$. Firstly, the tensor is reshaped to $H \times W \times (C * T)$ to combine the modality information into the channel dimension. Next, the typical channel attention mechanism (Hu, Shen, and Sun 2018) is applied to the reshaped features as shown in Fig. 5, and finally the attended feature tensor is reshaped back from $(H \times W \times (C * T))$ to $T \times H \times W \times C$ and then added with the original tensor to form a residual attention manner. Our experimental results show that CMA outperforms the naive 3D squeeze-excitation block and is more suitable for our framework.

Experiments and Results

Datasets and Metrics

We evaluate our *RD3D* on six popular public datasets having paired RGB and depth images, including: NJU2K (1,985 pairs) (Ju et al. 2014)), NLPR (1,000 pairs) (Peng et al. 2014), STERE (1,000 pairs) (Niu et al. 2012), DES (135 pairs, also called the RGBD135 dataset in some previous works) (Cheng et al. 2014), SIP (929 pairs) (Fan et al. 2020a) and DUTLF-D (1,200 pairs) (Piao et al. 2019). Following (Chen and Li 2018; Chen, Li, and Su 2019; Han et al. 2017), we use the same 1,485 pairs from NJU2K and 700 pairs from NLPR for training. The remaining pairs are used for testing. Specially, on the latest DUTLF-D dataset, we follow (Piao et al. 2019; Zhao et al. 2020; Piao et al. 2020; Li et al. 2020; Ji et al. 2020) to add additional 800 pairs from DUTLF-D for training and test on the remaining 400 pairs. In summary, our training set contains 2,185 paired RGB and depth images except when testing is conducted on DUTLF-D.

We use the newly proposed S-measure (S_α) (Fan et al. 2017) and E-measure (E_ϕ) (Fan et al. 2018b), as well as the generally agreed F-measure (F_β) (Borji et al. 2015) and Mean Absolute Error ($\mathcal{M}$) (Perazzi et al. 2012) as evaluation metrics for comparing performance of different models. These four metrics provide comprehensive and reliable evaluation results and have been adopted by many previous works. Following (Fu et al. 2020a), we report the maximum F-measure ($F_\beta^{\max}$) and maximum E-measure ($E_\phi^{\max}$) scores.

Implementation Details

3D ResNet We implement our 3D ResNet encoder based on the 2D ResNet (He et al. 2016). We replace all 2D kernels in the ResNet-50 with their 3D versions and the 3D kernel weights are initialized by the 2D weights pre-trained on ImageNet (Russakovsky et al. 2015) in a centralized initialization manner (Girdhar et al. 2018). We reduce the channel numbers of different side outputs to a fixed number 32 in the channel reduction (CR) modules.

Training and Testing Settings Our framework is implemented based on PyTorch (Paszke et al. 2019) on a work-

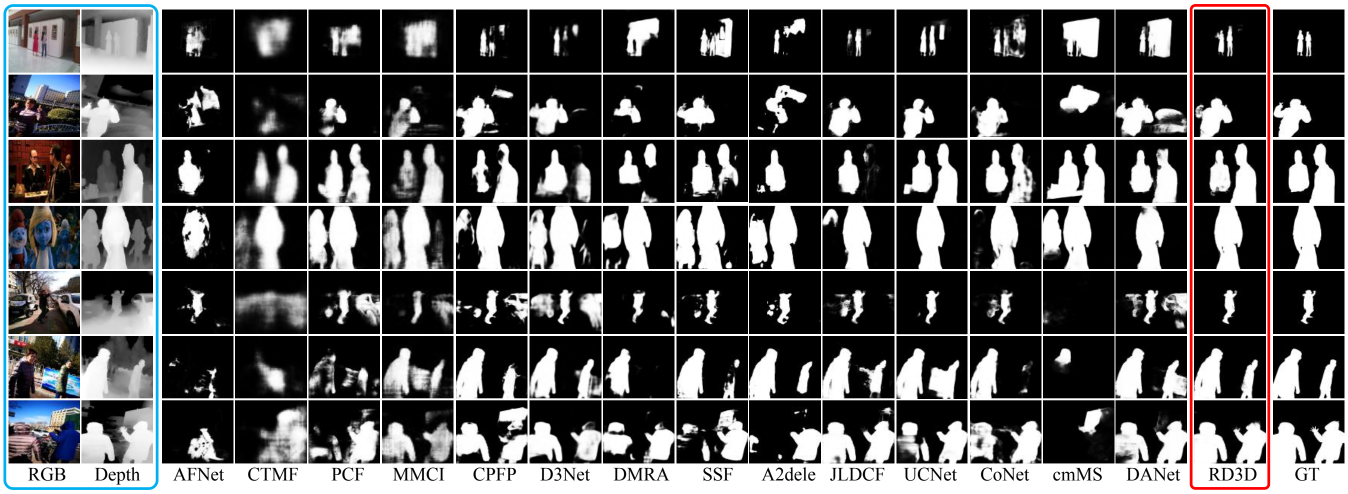


Figure 6: Qualitative comparisons of *RD3D* with state-of-the-art (SOTA) methods. “GT” indicates the ground truth.

Architecture	Speed (fps)	Size (MB)	NLPR (500 pairs)				NJU2K (300 pairs)				STERE (1000 pairs)				SIP (929 pairs)			
			$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$
Input Fusion	47.4	94.4	.919	.027	.901	.953	.904	.043	.904	.937	.892	.047	.885	.935	.876	.053	.879	.917
Two-stream	32.5	200.0	.929	.023	.918	.962	.913	.039	.911	.944	907	.040	.899	.941	.878	.052	.881	.923
Siamese	46.4	94.4	.927	.024	.917	.959	.915	.037	.913	.946	.904	.041	.898	.939	.867	.057	.867	.905
RD3D	**45.6**	**180.8**	**.930**	**.022**	**.919**	**.965**	**.916**	**.036**	**.914**	**.947**	**.911**	**.037**	**.906**	**.947**	**.885**	**.048**	**.889**	**.924**

Table 2: Comparisons of different backbone strategies on four large datasets. The results of our *RD3D* are highlighted in bold. Here F_β, E_ϕ mean $F_\beta^{\max}$ and $E_\phi^{\max}$, respectively, whose superscripts are omitted for the sake of space.

station with 4 NVIDIA 1080Ti GPUs. During training, we adopt the Adam optimizer with an initial learning rate of 0.0001, which is decayed by a cosine learning rate scheduler. The weight decay is set to 0.001. The data is first resized to $[352, 352]$ and then augmented by random horizontal flip and multi-scale transformation with the scale of {256, 352, 416}. We train for 100 epochs on 4 GPUs with the batch size equals to 10 per GPU, and the total training time is about 6 hours. The model after the last epoch is used for inference. Regarding the supervision, we calculate the typical binary cross-entropy loss. During testing, an image of arbitrary size is first resized to $[352, 352]$ and the predicted saliency map is resized back to its original size.

Comparisons with SOTAs

We compare *RD3D* with 14 SOTA deep RGB-D SOD models, including AFNet (Wang and Gong 2019), CTMF (Han et al. 2017), PCF (Chen and Li 2018), MMCI (Chen, Li, and Su 2019), CPFP (Zhao et al. 2019a), D3Net (Fan et al. 2020a), DMRA (Piao et al. 2019), SSF (Zhang et al. 2020c), A2dele (Piao et al. 2020), JL-DCF (Fu et al. 2020a), UCNet (Zhang et al. 2020b,a), CoNet (Ji et al. 2020), cmMS (Li et al. 2020) and DANet (Zhao et al. 2020). Quantitative results are shown in Table 1. It can be seen that compared with other methods, our results have notable improvement on the six datasets, advancing the best scores obtained by SOTA models by an average of 0.68%/0.50% on $S_\alpha/F_\beta^{\max}$. We show visualization results of *RD3D* and other methods in Fig. 6. In the global view, the detection of *RD3D* is more accurate. In the detailed view, *e.g.*, in the first row of Fig. 6, only *RD3D* can accurately identify the two people as the foreground. In general, the decent qualitative performance of *RD3D* is consistent with the quantitative analysis.

Ablation Studies: Backbone Strategies

To validate the pre-fusion in the backbone via 3D CNNs, we compare the four backbone strategies shown in Fig. 1 (a)-(d). Our method belongs to Fig. 1 (d), and we implement Input Fusion Network (Fig. 1 (c)), Two-stream Network (Fig. 1 (a)), and Siamese Network (Fig. 1 (d)) by switching the encoder part of *RD3D*. Note that the main difference lies in the way the encoder deals with multi-modal inputs. For fair comparison, we keep the decoder the same. We implement Input Fusion Network by first concatenating the RGB and depth images in the channel dimension and then fusing them by the first convolution layer in the 2D ResNet. Since the input shape is inconsistent with the original ResNet, we modify the first convolution layer and later repeat the encoder outputs in the temporal axis to enforce input $T = 2$ for the decoder. For Two-stream Network, we use two 2D ResNets to extract hierarchical features separately. Likewise, features are then concatenated in the temporal axis. The Siamese Network is implemented by a shared 2D ResNet for RGB and depth, while keeping other settings the same.

Table 2 shows experimental results on four large datasets including NLPR, NJU2K, STERE and SIP. As can be seen,

Model	Speed	Size	NLPR (500 pairs)				NJU2K (300 pairs)				STERE (1000 pairs)				SIP (929 pairs)			
	(fps)	(MB)	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$
DANet	32.0	106.7	.915	.028	.903	.953	.899	.045	.898	.935	.901	.043	.892	.937	.875	.054	.876	.918
JL-DCF	9.0	520.0	.925	.022	.916	.962	.903	.043	.903	.944	.905	.042	.901	.946	.879	.051	.885	.923
RD3D	**45.6**	**180.8**	**.930**	**.022**	**.919**	**.965**	**.916**	**.036**	**.914**	**.947**	**.911**	**.037**	**.906**	**.947**	**.885**	**.048**	**.889**	**.924**
Model-1	52.5	180.5	.913	.031	.894	.949	.906	.043	.898	.940	.897	.049	.884	.935	.873	.059	.867	.915
Model-2	50.2	180.5	.918	.028	.899	.949	.913	.040	.913	.944	.906	.042	.897	.940	.878	.053	.882	.919
Model-3	45.8	180.7	.921	.027	.904	.949	.914	.039	.913	.942	.907	.042	.897	.939	.866	.059	.864	.901
Model-4	40.4	219.1	.931	.022	.921	.965	.920	.034	.923	.952	.908	.039	.901	.944	.883	.048	.890	.924

Table 3: Ablation results on four large datasets, where F_β, E_ϕ mean $F_\beta^{\max}$ and $E_\phi^{\max}$. The results of our *RD3D* are in bold.

RD3D based on 3D CNNs outperforms the other three strategies by a notable margin. The Input Fusion Network performs worst though its model size is small, because multi-modality inputs are fused too naively, leading to insufficient extraction of multi-modal information. Besides, the Two-stream Network and Siamese Network are comparable to each other, but both are worse than our strategy. This clearly demonstrates the effectiveness of pre-fusion in the backbone through 3D convolutions. Regarding the model speed and size of our scheme, they are almost equal to those of the Two-stream Network, but our numbers are slightly better.

Ablation Studies: Other Modules

We take the full model of *RD3D* as the reference and conduct thorough ablation studies by replacing or removing the key components. The full version is denoted as RD3D (3D ResNet+CMA +RBPP), where "CMA" and "RBPP" refer to the usage of CMA modules and RBPP. Firstly, we construct a baseline "Model-1" (3D ResNet) by removing CMA modules and RBPP. Thus, the decoder of this model is just a plain 3D UNet decoder. Secondly, to validate the effectiveness of the rich back-projection paths (RBPP), we realize "Model-2" (3D ResNet+CMA) by removing all the back-projection paths. Thirdly, to demonstrate the benefit of channel-modality attention modules (CMA), we implement "Model-3" (3D ResNet+CA+RBPP), which replaces all CMA modules with naive squeeze-excitation channel attention modules (Hu, Shen, and Sun 2018), namely only channel attention is considered and during the squeeze operation, the global pooling is applied to the other three dimensions. Lastly, to investigate the proposed CMA modules, we also construct "Model-4" (3D ResNet+CMA*+RBPP), where "CMA*" means moving CMA from the decoder to the encoder stage. CMA modules are inserted into the ResNet backbone in a way as suggested by (Hu, Shen, and Sun 2018). Results of the above ablation studies are reported in Table 3, where two SOTA models DANet and JL-DCF are listed also. The following observations can be achieved.

Effectiveness of the Baseline Model. Without bells and whistles, the baseline model "Model-1" performs favorably against the two latest SOTA models DANet and JL-DCF, showing the potentials of using 3D convolutions for achieving effective cross-modality feature aggregation. Note that this baseline model consists of only basic 3D convolutions without any other augmentation.

Effectiveness of RBPP. Comparing between "Model-2" and the full *RD3D* in Table 3 shows that removing the RBPP leads to consistent performance degeneration. This implies that taking use of all information from higher-resolution levels is beneficial, especially to our framework where the back-projection paths contain rich multi-level modality-aware information.

Effectiveness of CMA. Comparing "Model-3" to *RD3D* in Table 3, one see that when the CMA modules are replaced, the performance drops, demonstrating that our channel-modality attention mechanism can enhance the final prediction and is probably more suitable for our fully 3D CNNs-based framework. Comparing "Model-2" to "Model-1", without RBPP, the improvement from adding CMA is still notable. This implies that combining RBPP and CMA is a reasonable and effective design, which results in substantial enhancement. In addition, "Model-4" achieves slightly better benchmark results than *RD3D*, showing that the proposed CMA modules can also work on the backbone. However, moving CMA to the encoder leads to slightly the higher computation and model size as in Table 3, because much more CMA modules have been deployed. Since using attention modules in the backbone is usually not adopted by the previous works, for fair comparison, we opt to deploy CMA in the decoder of *RD3D*.

Conclusion

We propose a novel RGB-D SOD framework called *RD3D*, which is based on 3D CNNs and conducts cross-modal feature fusion in a progressive manner. *RD3D* first utilizes 3D convolutions for pre-fusion between RGB and depth, and then conduct explicit fusion of modality-aware features by a 3D decoder augmented with rich back-projection paths and channel-modality attention modules. Extensive experiments on six benchmark datasets demonstrate that *RD3D*, which is the first fully 3D CNNs-based RGB-D SOD model, performs favorably against existing SOTA approaches. Detailed ablation studies and discussions validate the key components of *RD3D*. In the future, we hope *RD3D* could encourage more RGB-D SOD designs based on 3D CNNs.

Acknowledgments

This work was supported by the NSFC, under No. 61703077, 61773270, 61971005, the Fundamental Research Funds for the Central Universities No. YJ201755.

References

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2019. Voxelmorph: a learning framework for deformable medical image registration. *IEEE TIP* 38(8): 1788–1800.

Borji, A.; Cheng, M.-M.; Jiang, H.; and Li, J. 2015. Salient object detection: A benchmark. *IEEE TIP* 24(12): 5706–5722.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Chen, H.; and Li, Y. 2018. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *CVPR*, 3051–3060.

Chen, H.; Li, Y.; and Su, D. 2019. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition* 86: 376–385.

Chen, Q.; Fu, K.; Liu, Z.; Chen, G.; Du, H.; Qiu, B.; and Shao, L. 2020. EF-Net: A Novel Enhancement and Fusion Network for RGB-D Saliency Detection. *Pattern Recognition* 107740.

Cheng, Y.; Fu, H.; Wei, X.; Xiao, J.; and Cao, X. 2014. Depth enhanced saliency detection method. In *ICIMCS*, 23–27.

Fan, D.-P.; Cheng, M.-M.; Liu, J.-J.; Gao, S.-H.; Hou, Q.; and Borji, A. 2018a. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 186–202.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018b. Enhanced-alignment measure for binary foreground map evaluation. *IJCAI* 698–704.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020a. Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. *IEEE TNNLS* .

Fan, D.-P.; Wang, W.; Cheng, M.-M.; and Shen, J. 2019. Shifting more attention to video salient object detection. In *CVPR*, 8554–8564.

Fan, D.-P.; Zhai, Y.; Borji, A.; Yang, J.; and Shao, L. 2020b. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *ECCV*.

Fan, X.; Liu, Z.; and Sun, G. 2014. Salient region detection for stereoscopic images. In *ICDSP*, 454–458. IEEE.

Feichtenhofer, C. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *CVPR*, 203–213.

Feichtenhofer, C.; Pinz, A.; and Wildes, R. P. 2016. Spatiotemporal residual networks for video action recognition. CoRR abs/1611.02155 (2016). *arXiv preprint arXiv:1611.02155* .

Feng, D.; Barnes, N.; You, S.; and McCarthy, C. 2016. Local background enclosure for RGB-D salient object detection. In *CVPR*, 2343–2350.

Fu, K.; Fan, D.-P.; Ji, G.-P.; and Zhao, Q. 2020a. JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, 3052–3062.

Fu, K.; Fan, D.-P.; Ji, G.-P.; Zhao, Q.; Shen, J.; and Zhu, C. 2020b. Siamese network for rgb-d salient object detection and beyond. *arXiv preprint arXiv:2008.12134* .

Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; and Tran, D. 2018. Detect-and-track: Efficient pose estimation in videos. In *CVPR*, 350–359.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 6047–6056.

Guo, C.; and Zhang, L. 2009. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP* 19(1): 185–198.

Han, J.; Chen, H.; Liu, N.; Yan, C.; and Li, X. 2017. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE TCYB* 48(11): 3171–3183.

Han, J.; Ngan, K. N.; Li, M.; and Zhang, H.-J. 2005. Unsupervised extraction of visual attention objects in color images. *TCSVT* 16(1): 141–145.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *CVPR*, 3203–3212.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*, 7132–7141.

Huang, P.; Shen, C.-H.; and Hsiao, H.-F. 2018. Rgbd salient object detection using spatially coherent deep learning framework. In *ICDSP*, 1–5.

Itti, L. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE TIP* 13(10): 1304–1318.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE TPAMI* 35(1): 221–231.

Ji, W.; Li, J.; Zhang, M.; Piao, Y.; and Lu, H. 2020. Accurate rgb-d salient object detection via collaborative learning. *ECCV* .

Jiang, Y.; Zhou, T.; Ji, G.-P.; Fu, K.; Zhao, Q.; and Fan, D.-P. 2020. Light Field Salient Object Detection: A Review and Benchmark. *arXiv preprint arXiv:2010.04968* .

Ju, R.; Ge, L.; Geng, W.; Ren, T.; and Wu, G. 2014. Depth saliency based on anisotropic center-surround difference. In *ICIP*, 1115–1119.

Li, C.; Cong, R.; Piao, Y.; Xu, Q.; and Loy, C. C. 2020. Rgb-d salient object detection with cross-modality modulation and selection. In *ECCV*.

Li, G.; Liu, Z.; and Ling, H. 2020. ICNet: Information Conversion Network for RGB-D Based Salient Object Detection. *IEEE TIP* 29: 4873–4884.

Li, H.; Chen, G.; Li, G.; and Yu, Y. 2019. Motion guided attention for video salient object detection. In *ICCV*, 7274–7283.

Li, Q.; Zhou, Y.; and Yang, J. 2011. Saliency based image segmentation. In *ICMT*, 5068–5071.

Liu, N.; Zhang, N.; and Han, J. 2020. Learning Selective Self-Mutual Attention for RGB-D Saliency Detection. In *CVPR*, 13756–13765.

Liu, Z.; Shi, S.; Duan, Q.; Zhang, W.; and Zhao, P. 2019. Salient object detection for RGB-D image by single stream recurrent convolution neural network. *Neurocomputing* 363: 46–57.

Niu, Y.; Geng, Y.; Li, X.; and Liu, F. 2012. Leveraging stereopsis for saliency analysis. In *CVPR*, 454–461.

Pang, Y.; Zhang, L.; Zhao, X.; and Lu, H. 2020. Hierarchical Dynamic Filtering Network for RGB-D Salient Object Detection. *ECCV* .

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 8026–8037.

Peng, H.; Li, B.; Xiong, W.; Hu, W.; and Ji, R. 2014. Rgb-d salient object detection: a benchmark and algorithms. In *ECCV*, 92–109.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 733–740.

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, 7254–7263.

Piao, Y.; Rong, Z.; Zhang, M.; Ren, W.; and Lu, H. 2020. A2dele: Adaptive and Attentive Depth Distiller for Efficient RGB-D Salient Object Detection. In *CVPR*, 9060–9069.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3): 211–252.

Song, H.; Liu, Z.; Du, H.; Sun, G.; Le Meur, O.; and Ren, T. 2017. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE TIP* 26(9): 4204–4216.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497.

Wang, N.; and Gong, X. 2019. Adaptive fusion for RGB-D salient object detection. *IEEE Access* 7: 55277–55284.

Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H.; and Yang, R. 2019. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146* .

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*, 7794–7803.

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F.; Aliakbarian, S.; and Barnes, N. 2020a. Uncertainty Inspired RGB-D Saliency Detection. *arXiv preprint arXiv:2009.03075* .

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F. S.; Zhang, T.; and Barnes, N. 2020b. UC-Net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, 8582–8591.

Zhang, M.; Ren, W.; Piao, Y.; Rong, Z.; and Lu, H. 2020c. Select, Supplement and Focus for RGB-D Saliency Detection. In *CVPR*, 3472–3481.

Zhang, P.; Liu, W.; Wang, D.; Lei, Y.; Wang, H.; and Lu, H. 2020d. Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. *Pattern Recognition* 100: 107130.

Zhang, Z.; Lin, Z.; Xu, J.; Jin, W.; Lu, S.-P.; and Fan, D.-P. 2021. Bilateral attention network for rgb-d salient object detection. *IEEE TIP* .

Zhao, J.-X.; Cao, Y.; Fan, D.-P.; Cheng, M.-M.; Li, X.-Y.; and Zhang, L. 2019a. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *CVPR*, 3927–3936.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019b. EGNet: Edge guidance network for salient object detection. In *ICCV*, 8779–8788.

Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; and Zhang, L. 2020. A Single Stream Network for Robust and Real-time RGB-D Salient Object Detection. *ECCV* .

Zhou, T.; Fan, D.-P.; Cheng, M.-M.; Shen, J.; and Shao, L. 2021. RGB-D Salient Object Detection: A Survey. *CVM* .

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 4490–4499.