Signal Processing: Image Communication 104 (2022) 116674

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

AGRFNet: Two-stage cross-modal and multi-level attention gated recurrent fusion network for RGB-D saliency detection

Zhengyi Liu [a,*], Yuan Wang [a], Yacheng Tan [a], Wei Li [a], Yun Xiao [b]

[a] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, 230601, China

[b] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, 230601, China

ARTICLE INFO

Keywords:
Salient object detection
Gated recurrent unit
Attention mechanism
Cross-modal
Multi-level
RGB-D image

ABSTRACT

RGB-D saliency detection aims to identify the most attractive objects in a pair of color and depth images. However, most existing models adopt classic U-Net framework which progressively decodes two-stream features. In this paper, we decode the cross-modal and multi-level features in a unified unit, named Attention Gated Recurrent Unit (AGRU). It can reduce the influence of low-quality depth image, and retain more semantic features in the progressive fusion process. Specifically, the features of different modalities and different levels are organized as the sequential input, recurrently fed into AGRU which consists of reset gate, update gate and memory unit to be selectively fused and adaptively memorized based on attention mechanism. Further, two-stage AGRU serves as the decoder of RGB-D salient object detection network, named AGRFNet. Due to the recurrent nature, it achieves the best performance with the little parameters. In order to further improve the performance, three auxiliary modules are designed to better fuse semantic information, refine the features of the shallow layer and enhance the local detail. Extensive experiments on seven widely used benchmark datasets demonstrate that AGRFNet performs favorably against 18 state-of-the-art RGB-D SOD approaches.

1. Introduction

Salient object detection (SOD) aims to distinguish the most attractive object in a scene. It plays an important role in a wide range of computer vision tasks, such as image segmentation [1], tracking [2–4], retrieval [5], compression [6], edit [7], quality assessment [8], etc.

Recently with the widespread use of depth cameras, such as Kinect, iPhone XR and Huawei Mate10, SOD in RGB-D image has drawn a lot of attention because depth information is an important complementary modality for RGB image.

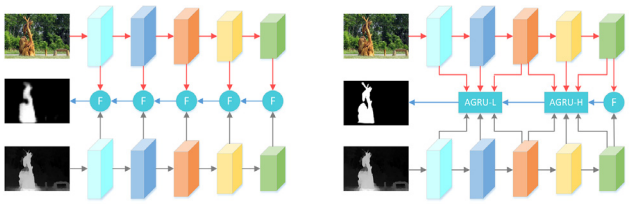
Classic U-Net framework [9] is widely adopted in the salient object detection area. Fig. 1(a) shows two-stream U-Net structure for RGB-D salient object detection. In the decoding process, the features of high layers are progressively transmitted and fused with the features of shallow layers from two streams by element-wise addition [10–12], concatenation [13–17], complementarity-aware fusion [18], fluid pyramid integration [19,20], dense aggregation [21–23], gated fusion [24–26], bi-directional fusion [27,28], residual fusion [29,30], cascaded partial fusion [31] or other fusion operations [32]. Many tricks are used to fuse cross-modal and multi-level features, but progressive fusion is their common spirit, in which the semantic information of deep layer may be gradually diluted in the progressive fusion process, and low-quality depth image also may make the fused result noisy.

Gated recurrent unit (GRU) is a particular type of recurrent neural networks (RNN). It is originally used for machine translation [33], and is later developed into a convolutional version for video representation [34]. Its gate mechanism and memory unit ensure that current output depends on history memory and current input, which can get the comprehensive judgment by long-range dependence. Inspired by GRU, an Attention GRU (AGRU) which introduces attention mechanism is proposed. It can selectively fuse cross-modal features by the gate structure, and retain the valid information and discard the irrelevant noise by memory structure, solving the issues about gradually diluted semantic information and heavy noise in depth image.

Further, the proposed AGRU is used to solve the problem of cross-modal and multi-level feature fusion in the salient object detection task in RGB-D images. But GRU is good at processing the sequential data with the same resolutions. Multi-level features extracted from the encoder should be upsampling to the same scale, which loss the accuracy necessarily. Therefore two-stage fusion strategy is proposed, which is shown in Fig. 1(b). The first AGRU of high layer (AGRU-H) fuses the features of last three layer, and the second AGRU of low layer (AGRU-L) fuses the features of the first three layers. Thus, cross-modal and multi-level features can be fused by two-stage AGRU to reduce the loss raised by excessively adjusting the resolutions of the features.

* Corresponding author.
E-mail addresses: liuzywen@ahu.edu.cn (Z. Liu), wangyuan.ahu@qq.com (Y. Wang), 1084043983@qq.com (Y. Tan), 1506375560@qq.com (W. Li), 280240406@qq.com (Y. Xiao).

https://doi.org/10.1016/j.image.2022.116674
Received 22 May 2021; Received in revised form 15 January 2022; Accepted 17 February 2022
Available online 4 March 2022
0923-5965/



(a) Two-stream U-Net framework (b) Our proposed two-stage AGRU framework

Fig. 1. Comparison between two-stream U-Net framework and our proposed two-stage AGRU framework.

Last, in order to further improve the performance, Correlation Fuse Module (CFM), Low-level Refinement Module (LRM) and Detail Enhance Module (DEM) are proposed. CFM fuses the cross-modal features of high layer from correlation attention aspect. LRM refines the features of low layer which will be fed into AGRU-L by the result of AGRU-H. DEM enhances the detail of result by residual learning.

Our main contributions can be summarized as follows:

- A cross-modal and multi-level attention gated recurrent fusion network is proposed for salient object detection in RGB-D image. It uses gated recurrent unit to fuse the cross-modal and multi-level features in the decoding process. Due to the resolution consistency requirement of input features in the recurrent structure, two-stage strategy is proposed to reduce the loss raised by excessive upsampling operation.
- GRU uses convolution operation as gate, which has the insufficient fusion effect. Attention GRU (AGRU) is proposed to enhance the cross-modal fusion or multi-level fusion in a unified structure based on attention mechanism. It can selectively combine color and depth features in the cross-modal fusion process, and adaptively remember the optimal fusion result in multi-level fusion process.
- In order to further improve the performance of our network, three modules (CFM, LRM, DEM) are proposed. They are designed to fuse the high-layer semantic information, refine the features of the shallow layer and enhance the local details.
- The proposed network is validated on seven publicly datasets, including NLPR [35], NJU2K [36], STERE [37], DES [38], SIP [39], DUT [40] and RedWeb-S [41]. The experimental results show that our network exhibits the excellent performance on these RGB-D saliency datasets.

2. Related work

2.1. RGB-D saliency detection

Salient object detection can detect the prominent objects in the RGB image [42], RGB-D image pairs [43–45], co-saliency image group [46–48], high-resolution image [49,50], light-field image [51,52], etc. Our research focuses on RGB-D SOD.

The fusion between two modalities is the important task of saliency detection for RGB-D images. Early fusion [30,53–55], middle fusion and late fusion [56–58] are three classic frameworks. The middle fusion widely used mainly adopts the classic two-stream U-Net framework in Fig. 1(a) to achieve the combination of different modal data, for example, PCFN [18], PGAR [29], ICNet [15], CMMS [32], ASIF-Net [17] etc. Depth induced network which uses depth information to improve RGB stream is another way, for example, PDNet [59], CFPF [19], A2dele [60], MobileSal [61], HDFNet [12], etc.

In order to solve the issue of low-quality depth map filtering, DCF [62] calibrates the latent bias in the original depth maps in advance, D3Net [39] uses gate mechanism to filter the poor depth map, EF-Net [63] enhances the depth maps by color hint map, DQSD [64] integrates a depth quality aware subnet into the classic bi-stream structure, assigning the weight of depth feature before conducting the selective RGB-D fusion. In addition, CoNet [65], DASNet [66], SSDP [67] and MobileSal [61] introduce depth estimation to improve the performance of SOD.

The past one year, the attention mechanism and edge guidance are two research hotspots in RGB-D saliency detection. SSF [68] adopts attention mechanism to enhance the feature representation, and introduces edge information to improve the performance. S2MA [14,41] proposes mutual attention mechanism which fuses multi-modal features of high layer inspired by non-local module [69]. CMMS [32] employs a saliency-guided position-edge attention (sg-PEA) module to focus more on saliency-related regions. cmSalGAN [70] incorporates edge information throughout cross-modal adversarial learning. FANet [71] introduces K-nearest neighbor graph neural networks and non-local module to obtain more discriminated geometric and appearance information.

There are other solutions. DCMF [72] embeds multi-modal image pairs into structural context space and content space to realize the cross-modal disentanglement and reconstruction. DRLF [73] replaces each sub-channel of RGB with depth information to generate DGB, RDB and RGD subnetwork, and conducts triple-stream fusion to achieve saliency detection. UCNet [74] learns the distribution of saliency maps rather than a single prediction based on a conditional variational autoencoder. RD3D [55] addresses RGB-D SOD through 3D convolutional neural networks. VST [44] designs Cross Modality Transformer to fuse color and depth features based on T2T-ViT [75] backbone network. CMINet [76] adopts mutual information minimization as a regularizer to reduce the redundancy between two modalities. SP-Net [77] proposes a shared learning network besides two modality-specific networks to jointly generate individual and shared saliency maps. DSA2F [78] presents an automatic architecture search to perform multi-modal feature fusion. TrasformerSOD [42] adopts swin transformer as backbone, and introduces the generator to produce saliency predictions with deep supervision and the confidence estimation based discriminator to estimate pixel-wise confidence map and achieve difficulty-aware learning.

Different from above mentioned methods, we propose to fuse cross-modal and multi-level features in a unified AGRU. The features are selectively fused by gate structure, and fusion result is optimized by retaining the valid information and discarding the irrelevant noise throughout the whole memory flow.

2.2. Decoding strategy

In the design of the decoder, element-wise addition and concatenation operation cannot better fuse cross-modal and multi-level features. PCFN [18] introduces complementarity-aware fusion (CA-Fuse) module, which involves cross-modal residual connections. It can capture the multi-level information and boost better cross-modal combination. CPFP [19] presents the fluid pyramid to integrate information in both multi-scale level and cross-modal level. It provides more interactions for cross-modal features in different scales. MCINet [20] adopts the similar method to conduct the decoding process. CMFS [21] densely accumulates features from the largest scale to the smallest scale. JL-DCF [22,23] augments the decoder with a dense connection [79] to promote the blending of depth and RGB features at various scales. FRDT [25] uses Gated Select Fusion Module (GSFM) to fuse cross-modal information and designs Adaptive Fusion Module (AFM) to fuse multi-level features. GFNet [24] proposes a gate fusion network (GFNet) with Res2Net architecture. CRACE [26] proposes a cross-attention context extraction module to serve as the decoder of an FPN-like network. MMNet [27] introduces Bi-directional Multi-scale Decoder (BMD) to aggregate multi-scale and comprehensive features from a top-down pathway and a bottom-up pathway. Zhang et al. [28] cascades both bottom-up and top-down feature aggregation paths in the decoding process. BiANet [30] proposes a bilateral attention module (BAM) to

comprise the dual complementary of foreground-first attention and background-first attention mechanisms. PGAR [29] proposes a Guided Residual Block (GRB) guided by initial saliency map, which receives RGB feature and depth feature alternately. BBS-Net [31,80] adopts two cascaded decoder stages inspired by [81]. Inspired by the T2T module [75], VST [44] designs reverse T2T module to achieve the decoding process.

Our proposed decoding strategy is to use a unified recurrent unit to fuse cross-modal and multi-level features. It is quite different from existing models which mainly adopt the progressive transition idea.

2.3. Gated recurrent unit

Gated Recurrent Unit (GRU) [33] is a particular type of Recurrent Neural Network (RNN). It is computationally less expensive and equally efficient as Long Short Term Memory (LSTM) [82]. It defines a recurrent hidden state whose activation at each time is dependent on that of the previous time. The reset gate r_t controls the degree to which the unit forgets the previously computed state. The update gate z_t determines the proportion of updates to hidden state. ConvGRU [34] is a convolutional version of GRU. So far, GRU has an outstanding performance in many natural language processing (NLP) tasks [83,84], video representation [34] and semantic scene completion [85], etc. Liu et al. [86] embed Gated Recurrent Unit (GRU) in each layer of the hourglass network to improve the long-range correlation of each side-output feature. Liu et al. [85] use multi-stage Gated Recurrent Fusion (GRF) module to adaptively select the multi-modal data and aggregate multi-level features with the same resolution. Fan et al. [87,88] use the similar method to dynamically combine the noisy and enhanced features in the speech recognition.

In the salient object detection, GRU has been widely used. Pahuja et al. [89] segment the image into patches, and then use a pair of encoder–decoder Conv-GRU [34] applied to patches and the whole image to improve the final segmentation map in a semantically coherent manner. Bardhan et al. [90,91] use Conv-GRU [34] which is called Contextual Unit (CTU) to refine each side-output feature, and empirically choose two time steps. Piao et al. [40] use ConvLSTM to iteratively learn the internal semantic relation of the fused feature. Shi et al. [92] use Bi-ConvLSTM to iteratively bridge the features in the encoder and decoder with the same resolutions. Wang et al. [93] use recurrent structure to achieve top-down and bottom-up inference by initial saliency map. Li et al. [94,95] design a Co-Attention Recurrent Unit to recurrently learn a robust group feature, which can handle the variation of co-object in appearance and the location across images. Li et al. [48] propose cycle-refinement module which utilizes ConvLSTMs to progressively update image embeddings and exchange information in a cycle manner. It is applied to the features on different levels individually to achieve the object co-segmentation.

In above methods, the gated recurrent structure is applied to one feature or the features of one layer, where the same resolution is their common characteristic. Different from it, our model decodes the cross-modal and multi-level features, which are viewed as sequential data. In order to solve the problem of resolution discrepancy, two-stage gated recurrent fusion framework is designed to reduce the loss from excessively adjusting the resolution.

3. The proposed method

3.1. Attention gated recurrent unit (AGRU)

Gated recurrent unit (GRU) is a popular RNN model which has been widely used in temporal related applications (e.g. video representation [34] and action recognition [96]). Its gate structure and memory unit can process the sequential data and generate the comprehensive results on the temporal dimension. In general, RGB-D image pairs can be encoded as cross-modal and multi-level features. We use GRU as decoder to selectively combine color and depth features in the cross-modal fusion process and adaptively remember the optimal fusion result in multi-level fusion process, which is shown in Fig. 1(b).

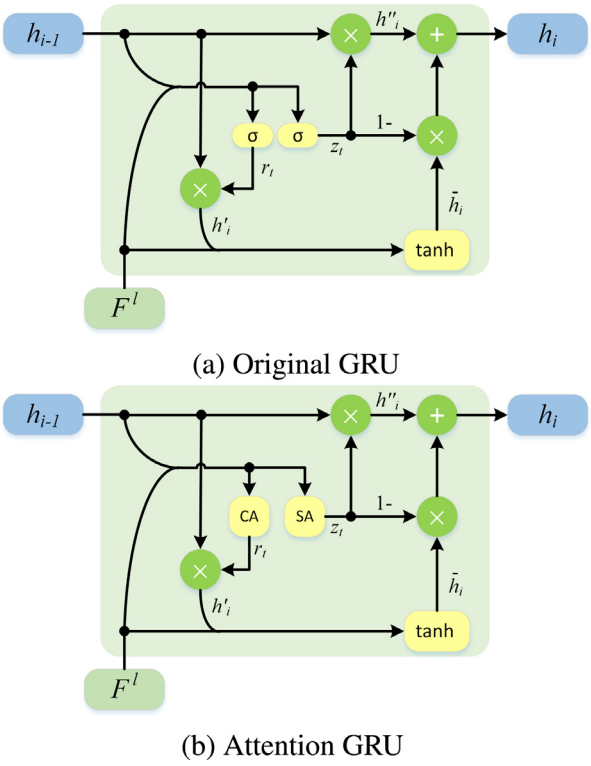


Fig. 2. The structures of GRU and Attention GRU (AGRU).

In each GRU, as shown in Fig. 2(a), the hidden state h_{i-1} represents the previous cross-modal and multi-level fusion result, and the current input F^l ($l \in \{c, d\}$) is RGB or depth features need to be fused.

The reset gate controls how much information in the past is retained. The update gate controls how much information in the past is discarded. The memory unit helps GRU to remember long-term information in the fusion process.

The reset gate r_t and the update gate z_t are defined as:

$$\begin{aligned} r_t &= \sigma(h_{i-1}, F^l) = sig(conv(cat(h_{i-1}, F^l))) \\ z_t &= \sigma(h_{i-1}, F^l) = sig(conv(cat(h_{i-1}, F^l))) \end{aligned} \tag{1}$$

where $cat(\cdot)$ denotes concatenation operation, $conv(\cdot)$ means the convolutional operation, and $sig(\cdot)$ is sigmoid function. The previous fusion result h_{i-1} and current input feature F^l are concatenated together, and then performed a convolutional operation, and last activated by a sigmoid function.

The reset gate r_t is used to reset the previous fusion result by:

$$h'_i = r_t \otimes h_{i-1} \tag{2}$$

where $\otimes$ denotes element-wise multiplication operation. The reset hidden state h'_i select the common salient parts in the previous fusion result h_{i-1} and current input F^l.

The update gate z_t is used to update the previous fusion result by:

$$h''_i = z_t \otimes h_{i-1} \tag{3}$$

It will discard the non-common salient regions in the previous fusion result and the current input by element-wise multiplication operation.

Then the current input F^l and reset hidden state h'_i are concatenated to form memory unit $\bar{h}_i$, which indicates the current memorized fusion result.

$$\bar{h}_i = tanh(conv(cat(F^l, h'_i))) \tag{4}$$

where $tanh(\cdot)$ is $tanh$ function.

Next, in order to supplement the salient part in the long-term information throughout the memory, the output hidden state h_i should add the non-common salient part deleted by update gate but prominent in the memory unit $\bar{h}_i$. The process can be described as:

$$h_i = h_i'' + (1 - z_t) \otimes \bar{h}_i \tag{5}$$

In a word, GRU first resets the historical hidden state based on the current input, combines the reset hidden state and current input to form memory unit, and then updates historical hidden state based on the current input, and last appends the salient part in the memory unit.

Furthermore, based on GRU, we propose Attention GRU (AGRU), as shown in Fig. 2(b). It replaces the reset gate and update gate with the channel attention CA and the spatial attention SA by:

$$r_t = CA(h_{i-1}, F^l) = sig(MLP(maxpool(cat(h_{i-1}, F^l)))) \tag{6}$$

$$z_t = SA(h_{i-1}, F^l) = sig(conv(maxpool_c(cat(h_{i-1}, F^l)))) \tag{7}$$

where $maxpool\ (\cdot)$ means the global max pooling operation, $MLP\ (\cdot)$ indicates a 2-layer perceptron, $maxpool_c\ (\cdot)$ denote the global max pooling operation for each point in the feature map along the channel axis. In the channel attention CA, the previous fusion result h_{i-1} and current input feature F^l are concatenated together, and then performed a global max pooling operation and 2-layer perceptron, and last activated by a sigmoid function. In the spatial attention SA, the previous fusion result h_{i-1} and current input feature F^l are concatenated together, and then performed a global max pooling operation along channel direction and a convolution layer, and last activated by a sigmoid function. They will serve as attention weights to replace reset gate and update gate.

The comparison experiment of GRU and AGRU (Section 4.4) demonstrates that attention can obtain the better weight than original gate structure to effective fuse cross-modal and multi-layer features. Meanwhile, we adopt the strategy that the channel attention CA serves the reset gate and the spatial attention SA serves as the update gate according to the comparison experimental result.

3.2. Cross-modal and multi-level attention gated recurrent fusion

Based on above proposed AGRU, we propose two-stage cross-modal and multi-level attention gated recurrent fusion network (AGRFNet). Its overall framework is depicted in Fig. 3, which consists of two-stream encoder, two-stage gated recurrent fusion decoder, a correlation fuse module, a low-level refinement module and a detail enhance module. The details can be seen in the following sections.

3.2.1. Two-stream encoder

Two-stream encoder utilizes two VGG/Resnet backbones to extract color feature and depth feature from RGB-D image, respectively.

In order to enhance the representation ability of side-output features, Convolutional Block Attention Module (CBAM) [97] and a transition layer are adopted. CBAM is used to emphasize the significant visual representations and suppress the noise of irrelevant clutters in channel and spatial dimensions, respectively. The transition layer contains a 3×3 convolution and a ReLU activation function, which can adjust the number of channels of side-output features to the same size 64. The side-output features of two streams after CBAM and transition layer are represented as F_i^c and F_i^d (i = 1, …,5), respectively.

3.2.2. Two-stage gated recurrent fusion decoder

The side-output features F_i^c and F_i^d are cross-modal and multi-level features. Their modalities are different, and resolutions are different too. In the process of decoding the cross-modal and multi-level features, element-wise addition and concatenation operation are widely used. The features of different layers are overlaid one by one from high layer to low layer. The semantic information in the high layer is progressively diluted in the data flow transmission, and inevitable mixed the noise in the shallow layer.

Due to selective fusion and adaptive memory ability of AGRU, cross-modal and multi-level features are organized as a sequence, and recurrently fed into AGRU. The gate structure in AGRU enables the selective fusion of multi-modal features. The memory structure in AGRU ensures that the feature of the high layer is effectively memorized throughout the transition process, avoid the dilution by the noisy feature of the low layer. AGRU is reused to fuse RGB feature and depth feature of different modalities and different levels.

But AGRU as a kind of RNN is not suitable for fusing multi-level features with different resolutions. The features must be adjusted to the same scales, which losses the accuracy necessarily. Therefore two-stage AGRU is adopted to fuse the multi-level features, so as to reduce the loss raised by the upsampled features. The first AGRU (AGRU-H) fuses the features of 3,4,5 layer to generate h^{high}, and the second AGRU (AGRU-L) fuses the features of 1,2,3 layer to generate h^{low}. The detail can be described as follows.

Input sequence. The depth features and RGB features are organized as a sequence H and L, and then fed into AGRU-H and AGRU-L in turn, respectively.

To be specific, F_5^l and F_4^l are upsampled with the same scales as F_3^l, where $l \in \{c, d\}$. They are organized as the input sequence with the same resolution about $64 \times 44 \times 44$ for AGRU-H.

$$H = \{\bar{F}_5^d, \bar{F}_5^c, \bar{F}_4^d, \bar{F}_4^c, F_3^d, F_3^c\} \tag{8}$$

$$\bar{F}_4^l = up_2(F_4^l), l \in \{c, d\} \tag{9}$$

$$\bar{F}_5^l = up_4(F_5^l), l \in \{c, d\} \tag{10}$$

where $up_2\ (\cdot)$ means the 2×upsampling operation using bilinear interpolation, $up_4\ (\cdot)$ means the 4×upsampling operation using bilinear interpolation.

In the similar strategy, F_3^l is 2×upsampled, F_2^l and F_1^l retains unchanged. They are organized as the input sequence with the same resolution about $64 \times 88 \times 88$ for AGRU-L.

$$L = \{\bar{F}_3^d, \bar{F}_3^c, F_2^d, F_2^c, F_1^d, F_1^c\} \tag{11}$$

$$\bar{F}_3^l = up_2(F_3^l), l \in \{c, d\} \tag{12}$$

Note that the order of color and depth features in the sequence H or L uses the depth-first and color-second manner. The opposite order, i.e. the color-first and depth-second manner, can achieve the same result. Because these two kinds of features as recurrent input will be selectively fused with the historical hidden state in AGRU, and they are equal input for AGRU, their order is not important.

Initial hidden state. The initial hidden state of AGRU-H is the fused features F_5^{CFM} from the highest layers of RGB and depth streams. The detail can be seen Section. 3.3.1. The initial hidden state of AGRU-L is the output of hidden state h^{high} in AGRU-H followed by an upsampling layer, which is denoted as h^{up}.

$$h^{up} = up_2(h^{high}) \tag{13}$$

Recurrent fusion process. We use the unfolded AGRU-H (Fig. 4) as an example to demonstrate the recurrent fusion process. From the figure, we can see a unified unit AGRU is reused to fuse both cross-modal color and depth features (F_i^c and F_i^d) and multi-level features (F_{i+1}^c and F_i^c or F_{i+1}^d and F_i^d). Specifically, the fused high layer feature F_5^{CFM} is used as the initial hidden state, and the sequence $H = \{\bar{F}_5^d, \bar{F}_5^c, \bar{F}_4^d, \bar{F}_4^c, F_3^d, F_3^c\}$ serves as the input sequence. The fused high layer feature is progressively combined with depth feature and color feature of different layers by AGRU with shared weights. The gate structure with the attention mechanism filters out the noise of color or depth feature, and memory unit adaptive memorizes the optimal fusion result in the long-range dependency, avoiding the dilution of the semantic information of high layer throughout the top-down data transition flow.

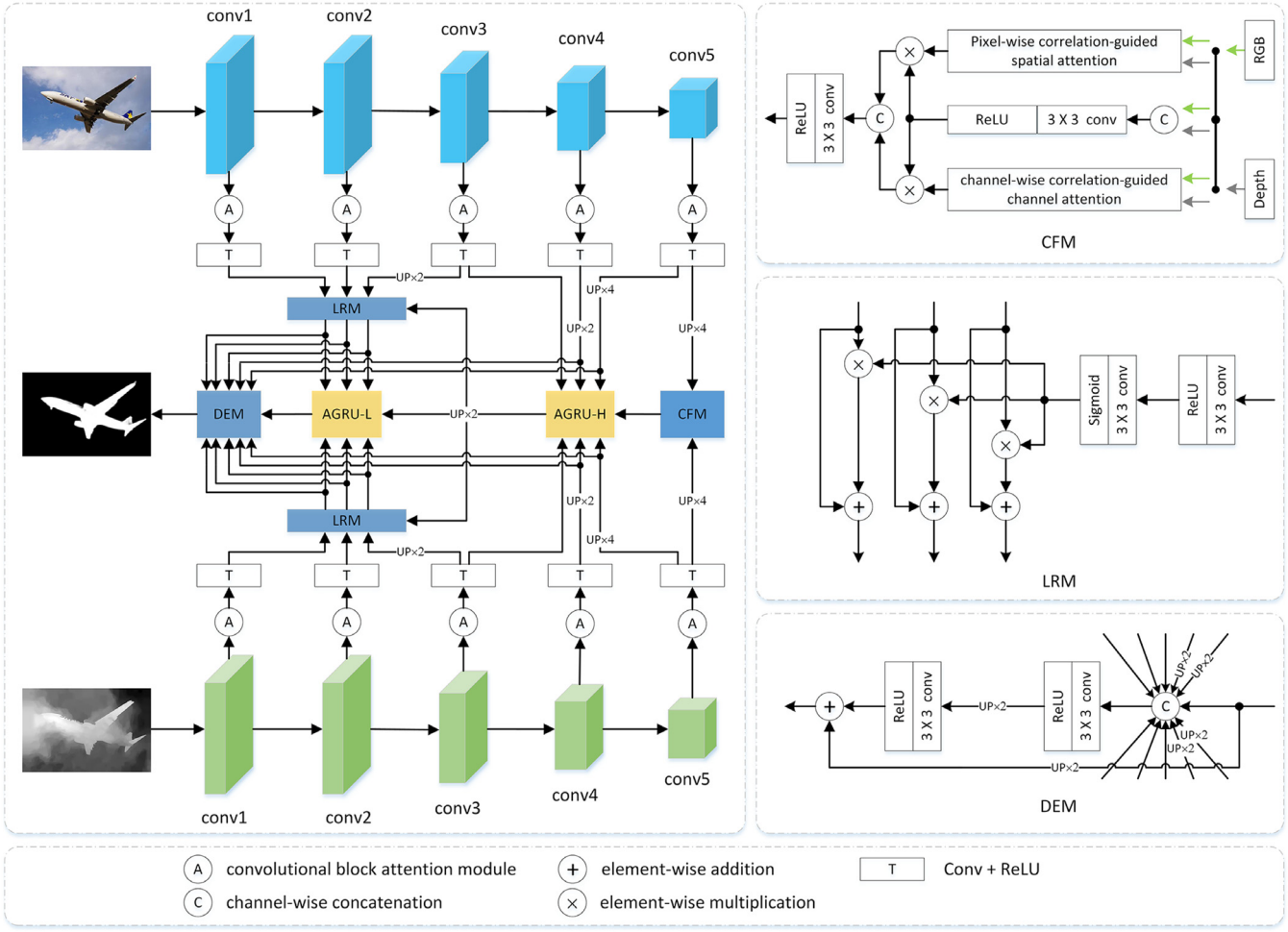


Fig. 3. The overall architecture of the Cross-modal and Multi-level Attention Gated Recurrent Fusion Network (AGRFNet).

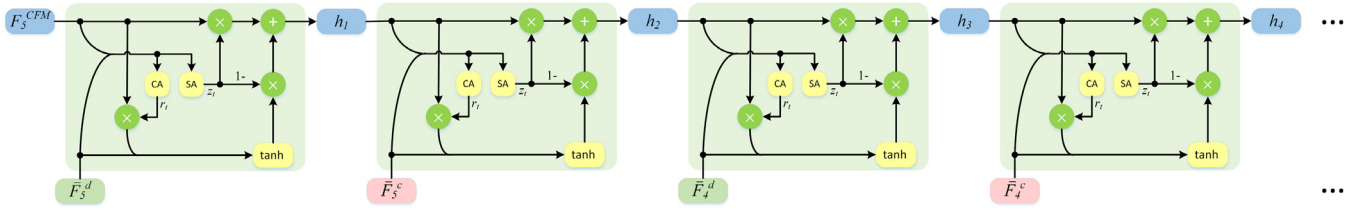


Fig. 4. The unfolded AGRU-H.

3.3. Three auxiliary modules

In order to improve the two-stage AGRU fusion performance, three auxiliary modules are designed. The details are shown as follows.

3.3.1. Correlation fuse module (CFM)

RGB and depth features are two different modalities for an image. The salient objects in the RGB feature should be prominent in the depth feature. The similarity map between RGB and depth feature can emphasize the common part. We introduce pixel-wise correlation-guided spatial attention [98] to learn a spatial descriptor that emphasizes informative position and suppresses the useless ones by taking the pixel-wise similarity maps as input. In the same way we also employ channel-wise correlation-guided channel attention [98]. They serve as the weight of cascaded features to generate more discriminative fusion feature which are used as initial hidden state of AGRU-H. CFM is shown in the top right of Fig. 3.

Specifically, we first cascade the 4×upsampled RGB high-level features $\bar{F}_5^c$ and depth high-level features $\bar{F}_5^d$ followed by a convolution layer with 3 × 3 kernel size and 64 channels to generate the fused feature F_5 by non-linear activation function ReLU. The process can be formulated as:

$$F_5 = ReLU(conv(cat(\bar{F}_5^c, \bar{F}_5^d))) \tag{14}$$

where $ReLU(\cdot)$ indicates the ReLU function.

Second, we use pixel-wise similarity maps between RGB features $\bar{F}_5^c$ and depth features $\bar{F}_5^d$ to learn spatial attention A_s, and use channel-wise correlation to learn channel attention A_c. The detail can be seen the Ref. [98]. The pixel-wise correlation-guided spatial attention and channel-wise correlation-guided channel attention can assign more weight on salient region commonly highlighted by RGB and depth features.

Last, we multiply the cascaded features F_5 with the spatial and channel attention maps to highlight the common salient regions, then

further cascade them to form F_5^{CFM}. The process can be formulated as:

$$F_5^{CFM} = ReLU(conv(cat(F_5 \otimes A_s, F_5 \otimes A_c))) \tag{15}$$

3.3.2. Low-level refinement module (LRM)

The output of AGRU-H has fused the cross-modal features of the high three layers. It can be used to refine the low-level features with some noises. LRM is shown in the center right of Fig. 3.

Specifically, the output of AGRU-H h^{high} is first upsampling to generate h^{up}. It is performed convolutional operation with 3×3 kernel and ReLU function to eliminate the aliasing effect caused by the upsampling operation and reduce the number of channels to 32. Next, it is performed a convolutional operation with 3×3 kernel and sigmoid function to generate initial saliency map S_H. The process can be formulated as:

$$S_H = sig(conv(ReLU(conv(h^{up})))) \tag{16}$$

Then the initial saliency S_H serves as the attention map to refine all the features in the input sequence $L = \{\bar{F}_3^d, \bar{F}_3^c, F_2^d, F_2^c, F_1^d, F_1^c\}$ of AGRU-L. Meanwhile, in order to preserve the original information, we also pile the initial features by a residual connection [10], and generate the refined low-level feature f', which can be described as:

$$f' = f \otimes S_H + f, f \in L \tag{17}$$

At last, all the features in L are updated by the high-level attention map and residual connection to form $L' = \{\bar{F}_3'^d, \bar{F}_3'^c, F_2'^d, F_2'^c, F_1'^d, F_1'^c\}$. They will serve as input sequence of AGRU-L.

3.3.3. Detail enhance module (DEM)

Two-stage AGRU starts from the top level and gradually uses the output of the higher level as prior knowledge to fuse the features of the lower level. Although the proposed recurrent fusion strategy adaptively fuses cross-modal and multi-level feature, it is not sufficient for detail recovery. Therefore, Detail Enhance Module (DEM) is designed to enhance the details of the features after two-stage AGRU. DEM is shown in the lower right of Fig. 3.

Specifically, we first cascade the output hidden state h^{low} of AGRU-L and all the input features of AGRU-H and AGRU-L to generate the cascaded feature F^{cat}. Note that F_3^c and F_3^d are removed in the concatenation process in order to avoid the repetitive use of the feature of the third layer. Then they are fed into two continuous blocks which consist of a convolution layer with 3×3 kernels and a ReLU layer. At last, a residual connection is added to make the final detailed result F^{ef} preserve the more information in h^{low}. The details are shown in the following formula:

$$F^{cat} = cat(h^{low}, f), f \in \{H - F_3^c - F_3^d\} \cup L' \tag{18}$$

$$F^{ef} = ReLU(conv(up_2(ReLU(conv(F^{cat}))))) + up_2(h^{low}) \tag{19}$$

The enhanced feature F^{ef} is used to generate the final saliency map by the following formula:

$$S_{final} = sig(conv(up_2(ReLU(conv(F^{ef}))))) \tag{20}$$

3.4. Loss function

As we all know, Binary Cross Entropy (BCE) loss is the most widely used loss function in SOD. However, the BCE loss only considers each pixel independently and does not consider the global structure of the image. In addition, BCE loss treats all pixels equally, which makes pixels in a cluttered area easy to be predicted error. Therefore, the proposed model uses pixel position aware loss L_{ppa} [99] between the saliency maps and the ground truth maps for end-to-end training.

It synthesizes local structure information to generate different weights for all pixels and introduces both pixel restriction $L_{wbce}(S, G)$ and global restriction $L_{wiou}(S, G)$, which can better guide the network learning and produce clear details.

The loss function is defined as:

$$L_{ppa}(S, G) = L_{wbce}(S, G) + L_{wiou}(S, G) \tag{21}$$

where S denotes predicted saliency map, G represents the ground truth map, L_{wbce} is a weighted binary cross entropy (wBCE) loss and L_{wious} denotes a weighted IoU (wIoU) loss. The detail can be seen the Ref. [99].

Finally, our total loss is defined as follows:

$$L = \lambda_1 L_{ppa}(S_{final}, G) + \lambda_2 L_{ppa}(S_H, G) \tag{22}$$

where S_{final} is the final predicted saliency map and S_H is the predicted saliency map output from AGRU-H. In this paper, we set $\lambda_1 = \lambda_2 = 1$.

4. Experiments

4.1. Datasets and evaluation metrics

4.1.1. Datasets

We evaluate the proposed method on seven challenging RGB-D SOD datasets. NLPR [35] includes 1000 images with single or multiple salient objects. NJU2K [36] consists of 1985 stereo image pairs and ground-truth maps with different objects, complex and challenging scenes. STERE [37] incorporates 1000 pairs of binocular images downloaded from the Internet. DES [38] has 135 indoor images collected by Microsoft Kinect. SIP [39] contains 929 high-resolution images of multiple salient persons. DUT [40] contains 1200 images captured by Lytro camera in real life scenes. ReDWeb-S [41] consists of 3179 images with high-quality depth maps from various real-world scenes. The dataset is split into a training set with 2179 RGB-D image pairs and a testing set with the remaining 1000 image pairs.

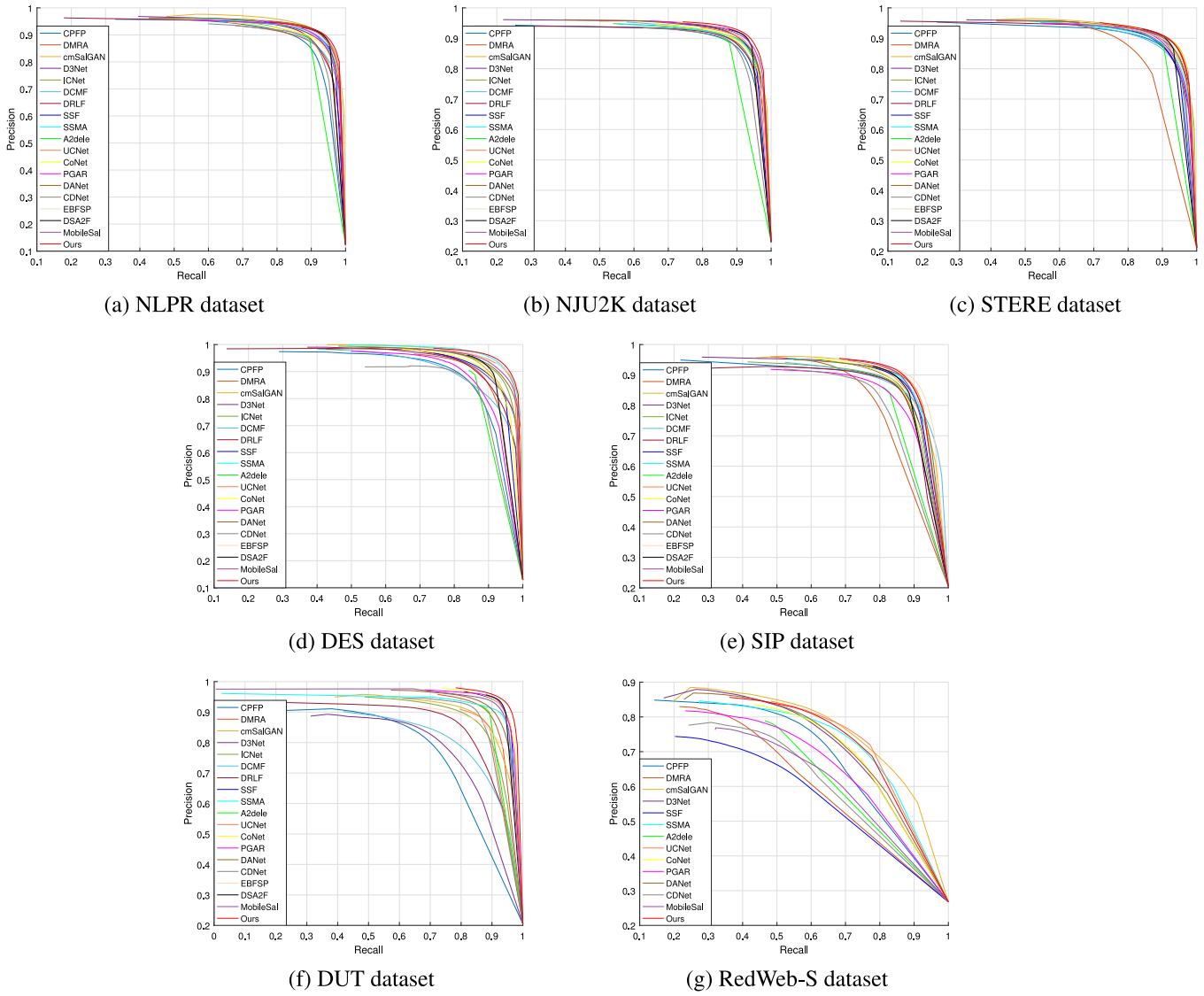
Training/Testing. For the sake of fair comparison, we use the same training dataset as in [29,39], which consists of 1485 images from the NJU2K dataset and 700 images from the NLPR dataset. The remaining images in the NJU2K and NLPR datasets, the whole datasets of STERE, DES, SIP and ReDWeb-S testing set are used for testing. In addition, on the DUT dataset, we follow the same protocols as in [32,40,54,60,65] to add additional 800 pairs from DUT for training and test on the remaining 400 pairs. In summary, our training set contains 2185 paired RGB and depth images, but when testing is conducted on DUT, our training set contains 2985 paired ones.

4.1.2. Evaluation metrics

We adopt five widely used metrics to evaluate the performance of our model and other state-of-the-art RGB-D SOD models, including the precision–recall (PR) curve [100], E-measure [101], S-measure [102], F-measure [103] and mean absolute error (MAE) [104]. Specifically, the PR curve plots precision and recall values by setting a series of thresholds on the saliency maps to get the binary masks and further comparing them with the ground truth maps. The E-measure simultaneously captures global statistics and local pixel matching information. The S-measure can evaluate both region-aware and object-aware structural similarity between saliency map and ground truth. The F-measure is the weighted harmonic mean of precision and recall, which can evaluate the overall performance. The MAE measures the average of the per-pixel absolute difference between the saliency maps and the ground truth maps. In our experiment, E-measure and F-measure adopts adaptive values.

4.2. Implementation details

During the training and testing phase, the input RGB and depth images are resized to 352×352. Multiple enhancement strategies



(a) NLPR dataset (b) NJU2K dataset (c) STERE dataset

(d) DES dataset (e) SIP dataset

(f) DUT dataset (g) RedWeb-S dataset

Fig. 5. P-R curves comparison of different models on seven datasets.

are used for all training images, i.e., random flipping, rotating and border clipping. Parameters of the backbone network are initialized with the pretrained parameters of ResNet-50 network [10]. The rest of parameters are initialized to PyTorch default settings. We employ the Adam optimizer [105] to train our network with a batch size of 5 and an initial learning rate 1e−4, and the learning rate will be divided by 10 every 60 epochs. Our model is implemented with PyTorch toolbox and trained on a machine with a single NVIDIA GTX 1080Ti GPU. The model converges within 200 epochs, which takes nearly 10 h. When testing, the proposed model runs at about 27 fps.

4.3. Comparisons with the-state-of-the-art

Our model is compared with 18 state-of-the-art RGBD SOD algorithms, including CPFP [19], DMRA [40], cmSalGAN [70], D3Net [39], ICNet [15], DCMF [72], DRLF [73], SSF [68], SSMA [14], A2dele [60], UCNet [74], CoNet [65], PGAR [29], DANet [54], CDNet [106], EBFSP [107], DSA2F [78] and MobileSal [61]. To ensure the fairness of the comparison results, the saliency maps of the evaluation are provided by the authors or generated by running pre-trained model.

Quantitative Evaluation. Fig. 5 shows the comparison results on PR curve. As can be clearly observed that our method achieves the good performance on seven datasets, especially on the NJU2K and DUT dataset. Furthermore, Table 1 shows the quantitative comparison results of four evaluation metrics. It can be seen that our method outperforms the other methods in terms of most metrics. Although some evaluation metrics are not the best, but they are better than those of most methods. The comparison results of PR curves and evaluation metrics demonstrate the effectiveness of AGRFNet which selectively combines color and depth features, adaptively remembers the optimal fusion result, and finally achieves the best performance by two-stage AGRU. Nevertheless, we also find that the evaluation value is a little worse than UCNet [74] on RedWebS. It indicates our method need to be improved to be more adaptive to real-world scenes.

Qualitative Evaluation. To make qualitative comparisons, we show some visual examples comparing our method with some top-ranking RGB-D saliency detection methods in Fig. 6. It can be observed that our method has the better detection results than other methods in some challenging cases: similar foreground and background ($1st$–$2nd$ rows), complex scene ($3rd$–$4th$ rows), low quality depth map ($5th$–$6th$ rows), small object ($7th$–$8th$ rows) and multiple objects ($9th$–$10th$ rows). In the $1st$–$4th$ rows, our model can successfully segment the object with the complete and clear outlines in the extremely similar or complex background. In the $5th$–$6th$ rows, our model can filter the noise of depth image. In the $7th$–$8th$ rows, our model can adaptive to small object. In the $9th$–$10th$ rows, multiple objects are all detected with our model,

Table 1
S-measure, adaptive F-measure, adaptive E-measure, MAE comparisons with different models. "–" means that the method does not release test results for this dataset. Red/Blue indicates the 1st/2nd result.

Datasets	Metric	CPFP CVPR19	DMRA ICCV19	cmSalGAN TMM20	D3Net TNNLS20	ICNet TIP20	DCMF TIP20	DRLF TIP20	SSF CVPR20	SSMA CVPR20	A2dele CVPR20	UCNet CVPR20	CoNet ECCV20	PGAR ECCV20	DANet ECCV20	CDNet TIP21	EBFSP TMM21	DSA2F CVPR21	MobileSal TPAMI21	AGRFNet Ours
NLPR	$S_\alpha\uparrow$	.888	.899	.922	.912	.923	.900	.903	.914	.915	.896	.920	.908	.918	.920	.902	.915	.918	.920	.923
	$F_\beta\uparrow$	.823	.854	.863	.861	.870	.839	.843	.875	.853	.878	.890	.846	.871	.875	.848	.897	.892	.877	.894
	$E_\xi\uparrow$	.924	.941	.947	.944	.944	.933	.936	.949	.938	.945	.953	.934	.948	.951	.935	.952	.950	.951	.951
	MAE↓	.036	.031	.027	.030	.028	.035	.032	.026	.030	.028	.025	.031	.028	.027	.032	.026	.024	.025	.025
NJU2K	$S_\alpha\uparrow$	.878	.886	.903	.901	.894	.889	.886	.899	.894	.869	.897	.895	.906	.899	.885	.903	.904	.905	.913
	$F_\beta\uparrow$	.837	.872	.874	.865	.868	.859	.849	.886	.865	.874	.889	.872	.883	.871	.866	.894	.898	.894	.903
	$E_\xi\uparrow$	.895	.908	.907	.914	.905	.897	.901	.913	.896	.897	.903	.912	.914	.908	.911	.907	.922	.914	.921
	MAE↓	.053	.051	.046	.046	.052	.052	.055	.043	.053	.051	.043	.046	.045	.045	.048	.039	.039	.040	.035
STERE	$S_\alpha\uparrow$	.879	.835	.896	.899	.903	.883	.888	.887	.890	.878	.903	.905	.903	.901	.896	.900	.897	.903	.903
	$F_\beta\uparrow$	.830	.845	.863	.859	.865	.841	.845	.867	.855	.874	.885	.884	.872	.868	.873	.870	.893	.879	.887
	$E_\xi\uparrow$	.903	.900	.914	.920	.915	.904	.915	.921	.907	.915	.922	.927	.917	.921	.922	.912	.927	.916	.927
	MAE↓	.051	.066	.050	.046	.045	.054	.050	.046	.051	.044	.039	.037	.044	.043	.042	.045	.039	.040	.038
SIP	$S_\alpha\uparrow$	.850	.806	.865	.860	.854	.859	.850	.868	.872	.826	.875	.858	.875	.875	.823	.885	.862	.873	.878
	$F_\beta\uparrow$	.819	.819	.849	.835	.836	.819	.813	.851	.854	.825	.868	.842	.848	.855	.805	.869	.865	.864	.876
	$E_\xi\uparrow$	.899	.863	.902	.902	.899	.898	.891	.911	.911	.892	.913	.909	.908	.914	.880	.917	.908	.914	.919
	MAE↓	.064	.085	.064	.063	.069	.068	.071	.056	.057	.070	.051	.063	.059	.054	.076	.049	.057	.053	.051
DES	$S_\alpha\uparrow$	.872	.901	.913	.898	.920	.877	.895	.905	.941	.885	.933	.911	.894	.924	.875	.937	.916	.929	.941
	$F_\beta\uparrow$	.829	.857	.869	.870	.889	.820	.868	.876	.906	.865	.917	.861	.870	.899	.839	.913	.901	.910	.923
	$E_\xi\uparrow$	.927	.945	.949	.951	.959	.923	.954	.948	.974	.922	.974	.945	.935	.968	.921	.974	.955	.973	.977
	MAE↓	.037	.029	.028	.031	.027	.040	.030	.025	.021	.028	.018	.027	.032	.023	.034	.018	.023	.021	.016
DUT	$S_\alpha\uparrow$	.749	.889	.867	.775	.852	.798	.826	.916	.903	.886	.864	.919	.920	.899	.880	.858	.921	.895	.928
	$F_\beta\uparrow$	.736	.884	.844	.756	.830	.750	.803	.914	.866	.890	.856	.909	.914	.888	.874	.842	.926	.908	.923
	$E_\xi\uparrow$	.815	.927	.897	.847	.897	.848	.870	.946	.921	.924	.903	.948	.944	.934	.918	.890	.950	.936	.948
	MAE↓	.100	.048	.067	.097	.072	.104	.080	.034	.044	.043	.056	.033	.035	.043	.048	.067	.030	.045	.029
RedWeb-S	$S_\alpha\uparrow$	.685	.592	.712	.689	–	–	–	.595	.711	.641	.713	.696	.656	.693	.635	–	–	.636	.705
	$F_\beta\uparrow$	.636	.576	.697	.664	–	–	–	.559	.687	.614	.711	.688	.642	.684	.601	–	–	.597	.703
	$E_\xi\uparrow$	.725	.671	.758	.742	–	–	–	.684	.747	.659	.758	.762	.729	.753	.691	–	–	.697	.756
	MAE↓	.142	.188	.139	.149	–	–	–	.189	.139	.160	.130	.147	.161	.142	.169	–	–	.175	.133

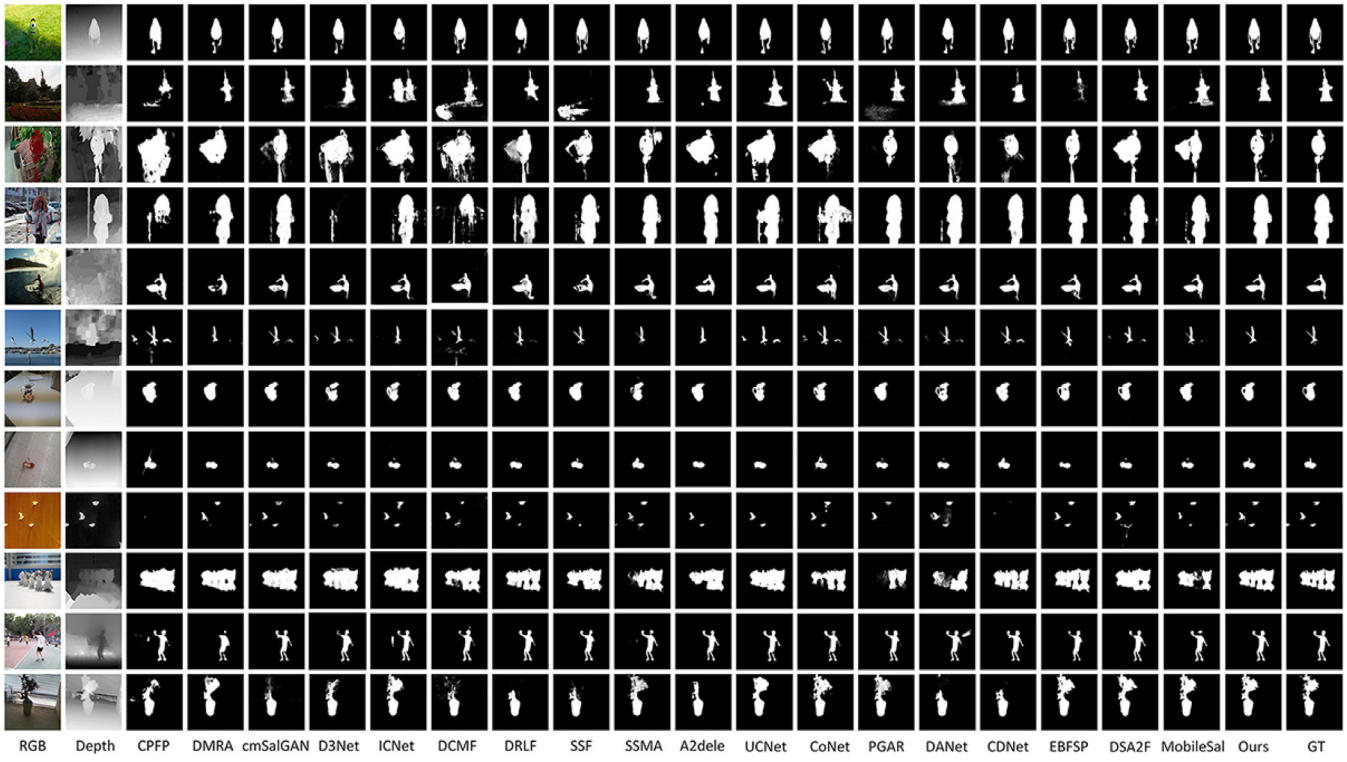


Fig. 6. Visual comparison results with other state-of-the-art models.

while some other models have the detection loss in numbers. The performances in these challenging cases verify the effectiveness of the designed AGRU and some auxiliary modules. In addition, our approach can produce more fine-grained details as highlighted in the salient region ($11th$–$12th$ rows). It is attributed to the designed two-stage AGRU decoding process.

Furthermore, to clearly show the contribution of AGRU on progressively decoding the cross-modal and multi-level features, the visualization comparison of U-Net decoding process and recurrent AGRU decoding process is shown in Fig. 7. From figure, we can observe that our features of high layer are clearer than that of U-Net framework, and the noise is better eliminated by our AGRU decoding. Finally, the better saliency map with sharper boundary is generated.

4.4. Ablation studies

We conduct ablation studies on NLPR, NJU2K and STERE datasets to investigate the contributions of different mechanisms in the proposed method.

The effectiveness of AGRU. To verify the effectiveness of AGRU, we conduct the ablation experiments whose results are shown in Table 2. The first model uses convolution and sigmoid activation function (denoted as σ) as reset and update gate. The last four models use four combinations of spatial attention (SA) and channel attention (CA) in reset and update gate. From the results, we can observe that the last four models are all better than the first model. It verifies that attention reset gate and update gate is better than original reset gate and update gate. Furthermore, the last four models have no obvious difference. The combination of CA and SA as reset gate and update gate is a little bit better. Therefore, the last model is adopted.

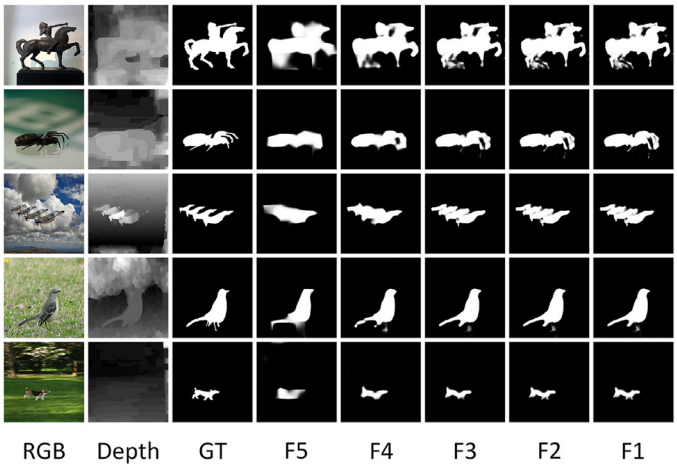
The effectiveness of two-stage AGRU decoder. The baseline model used here adopts the same framework as Fig. 1(a). It adopts two-stream U-Net structure and decodes the cross-modal and multi-level features by progressively fusion with element-wise addition operation. Its performance is illustrated in Table 3 No. 1. Based on the baseline, we adds CBAM modules to extract the attentive cross-modal and multi-level features, the result is shown in Table 3 No. 2. We find that CBAM module has the good performance in improving the feature representation.

In order to verify the effectiveness of two-stage AGRU decoder, ablation experiments are further conducted.

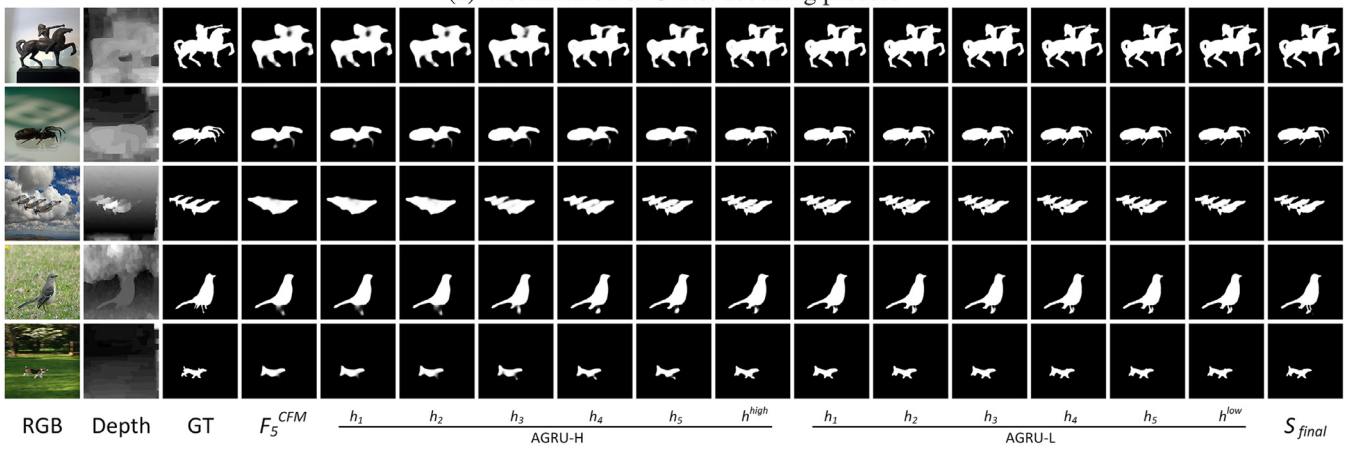
Table 3 No. 3 denotes the model which adopts one AGRU to fuse all the attentive features extracted from two-stream encoder. In order to use the recurrent structure, all the features must be adjusted to the same resolutions. It will loss the accuracy. Compared with No. 2, the performance of the variant No. 3 declines a bit. But the application of two-stage strategy significantly improves the performance, which is shown in Table 3 No. 4.

In addition, we also discuss the parameter number of the decoder among No. 2, No. 3 and No. 4, which is shown in Table 3. From the experimental result, we can find that due to the recurrent nature of AGRU, the number of parameters are significantly reduced compared with progressive fusion with element-wise addition operation. Two-stage AGRU achieves the best performance with the relative less parameters.

The effectiveness of three auxiliary modules. Based on the model No. 4, which is our two-stage AGRU framework, we gradually add auxiliary modules. These modules are correlation fuse module (CFM), low-level refinement module (LRM) and detail enhance module (DEM). In Table 3 No. 5, by applying CFM, the performance is boosted. It benefits from optimizing the cross-modal semantic information of high layer by pixel-wise correlation-guided spatial attention and channel-wise correlation-guided channel attention. In Table 3 No. 6, by applying LRM, the performance is enhanced further. It benefits from refining the contour detail of the features in the shallow layer by optimized cross-modal semantic information of high layer. In Table 3 No. 7, by applying DEM, the performance is improved again. It benefits from enhancing the local information by effective combination among the result of AGRU and all of side-output features. We can see that three auxiliary modules improve the performance gradually and obviously, verifying the effectiveness of our proposed modules.



(a) Visualization of U-Net decoding process



(b) Visualization of AGRU decoding process

Fig. 7. Visualization comparison of the decoding process.

Table 2
Ablation experiments of reset gate and update gate. The best result is in bold.

Variant	NLPR				NJU2K				STERE			
	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β↑	E_ξ ↑	MAE↓	S↑	F_β↑	E_ξ ↑	MAE↓
$\sigma+\sigma$	.922	.892	.949	.025	.911	.898	.910	.038	.902	.883	.923	.039
SA+SA	.923	**.898**	**.956**	**.024**	**.913**	**.903**	.912	.036	.902	.885	.925	.039
CA+CA	.919	.894	.953	.025	.912	.901	.916	.036	.901	.883	**.927**	.039
SA+CA	**.926**	**.898**	.955	**.024**	.912	.900	.914	.037	.902	.884	.926	.039
CA+SA (Ours)	.923	.894	.951	.025	**.913**	**.903**	**.921**	**.035**	**.903**	**.887**	**.927**	**.038**

Table 3
Ablation experiments of different components. The best result is in bold.

Variant	Candidate							Decoder parameter	NLPR				NJU2K				STERE			
	Baseline	CBAM	AGRU	2AGRU	CFM	LRM	DEM	(MB)	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β↑	E_ξ ↑	MAE↓	S↑	F_β↑	E_ξ ↑	MAE↓
No. 1	✓							–	.899	.856	.937	.033	.891	.862	.890	.049	.875	.840	.895	.054
No. 2		✓						24.9	.901	.861	.940	.032	.895	.870	.897	.047	.880	.846	.901	.053
No. 3		✓	✓					5.8	.899	.858	.938	.033	.893	.872	.899	.046	.877	.845	.904	.053
No. 4		✓		✓				6.6	.908	.867	.943	.030	.900	.883	.908	.043	.886	.862	.912	.048
No. 5		✓		✓	✓			–	.913	.879	.946	.028	.908	.891	.914	.039	.895	.875	.920	.044
No. 6		✓		✓	✓	✓		–	.919	.889	.949	.026	.910	.898	.919	.037	.899	.883	.924	.041
No. 7		✓		✓	✓	✓	✓	–	**.923**	**.894**	**.951**	**.025**	**.913**	**.903**	**.921**	**.035**	**.903**	**.887**	**.927**	**.038**

5. Conclusions

We propose an attention gated recurrent unit to fuse cross-modal and multi-level features in a unified recurrent structure. It introduces attention mechanism to design gate and memory unit, making it better retain useful information and discard irrelevant information. AGRU is used to decode the features of different modalities and different levels from RGB-D saliency detection encoder. In order to avoid the loss raised by upsampling operation, two-stage strategy is proposed. In addition, three auxiliary modules are designed to further improve the whole performance. Experimental results demonstrate the effectiveness of all the designs and strategies. In the future, we will use one AGRU in

decoding process, rather than two-stage AGRU, but more refinement operation, for example, edge supplement or something others need to be considered.

CRediT authorship contribution statement

Zhengyi Liu: Writing – original draft, Methodology, Supervision. **Yuan Wang:** Methodology, Visualization, Software, Validation. **Yacheng Tan:** Writing – review & editing. **Wei Li:** Writing– review & editing. **Yun Xiao:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62006002), Natural Science Foundation of Anhui Province, China (1908085MF182) and Science Research Project for Graduate Student of Anhui Province, China (YJS20210047).

References

[1] M. Donoser, M. Urschler, M. Hirzer, H. Bischof, Saliency driven total variation segmentation, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 817–824.
[2] C. Ma, Z. Miao, X.-P. Zhang, M. Li, A saliency prior context model for real-time object tracking, IEEE Trans. Multimed. 19 (11) (2017) 2415–2424.
[3] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: International Conference on Machine Learning, 2015, pp. 597–606.
[4] P. Zhang, W. Liu, D. Wang, Y. Lei, H. Wang, H. Lu, Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps, Pattern Recognit. 100 (2020) 107130.
[5] Y. Gao, M. Shi, D. Tao, C. Xu, Database saliency for fast image retrieval, IEEE Trans. Multimed. 17 (3) (2015) 359–369.
[6] Q.-G. Ji, Z.-D. Fang, Z.-H. Xie, Z.-M. Lu, Video abstraction based on the visual attention model and online clustering, Signal Process., Image Commun. 28 (3) (2013) 241–253.
[7] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1531–1544.
[8] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, H. Sun, Optimizing multistage discriminative dictionaries for blind image quality assessment, IEEE Trans. Multimed. 20 (8) (2017) 2035–2048.
[9] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[11] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, S. Kwong, Dense attention fluid network for salient object detection in optical remote sensing images, IEEE Trans. Image Process. 30 (2021) 1305–1317.
[12] Y. Pang, L. Zhang, X. Zhao, H. Lu, Hierarchical dynamic filtering network for RGB-D salient object detection, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, Springer, 2020, pp. 235–252.
[13] P.O. Pinheiro, T.-Y. Lin, R. Collobert, P. Dollár, Learning to refine object segments, in: European Conference on Computer Vision, Springer, 2016, pp. 75–91.
[14] N. Liu, N. Zhang, J. Han, Learning selective self-mutual attention for RGB-D saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13756–13765.
[15] G. Li, Z. Liu, H. Ling, ICNet: Information conversion network for RGB-D based salient object detection, IEEE Trans. Image Process. 29 (2020) 4873–4884.
[16] G. Li, Z. Liu, L. Ye, Y. Wang, H. Ling, Cross-modal weighting network for RGB-d salient object detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2020, pp. 665–681.
[17] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, Q. Huang, ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection, IEEE Trans. Cybern. 51 (1) (2020) 88–100.
[18] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060.
[19] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, L. Zhang, Contrast prior and fluid pyramid integration for RGBD salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3927–3936.
[20] Z. Huang, H.-X. Chen, T. Zhou, Y.-Z. Yang, B.-Y. Liu, Multi-level cross-modal interaction network for RGB-D salient object detection, Neurocomputing 452 (2021) 200–211.
[21] Y. Niu, G. Long, W. Liu, W. Guo, S. He, Boundary-aware RGBD salient object detection with cross-modal feature sampling, IEEE Trans. Image Process. 29 (2020) 9496–9507.
[22] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, C. Zhu, Siamese network for RGB-D salient object detection and beyond, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1–18.
[23] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, JL-DCF: JOint learning and densely-cooperative fusion framework for RGB-D salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3052–3062.
[24] W. Zhou, Y. Chen, C. Liu, L. Yu, GFNet: Gate fusion network with Res2Net for detecting salient objects in RGB-D images, IEEE Signal Process. Lett. 27 (2020) 800–804.
[25] M. Zhang, Y. Zhang, Y. Piao, B. Hu, H. Lu, Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4107–4115.
[26] P. Peng, Y.-J. Li, A unified structure for efficient RGB and RGB-D salient object detection, 2020, arXiv preprint arXiv:2012.00437.
[27] G. Liao, W. Gao, Q. Jiang, R. Wang, G. 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, 2020, pp. 2436–2444.
[28] Y.-f. Zhang, J. Zheng, L. Li, N. Liu, W. Jia, X. Fan, C. Xu, X. He, Rethinking feature aggregation for deep RGB-D salient object detection, Neurocomputing 423 (2021) 463–473.
[29] S. Chen, Y. Fu, Progressively guided alternate refinement network for RGB-D salient object detection, in: European Conference on Computer Vision, Springer, 2020, pp. 520–538.
[30] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, D.-P. Fan, Bilateral attention network for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 1949–1961.
[31] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, L. Shao, BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network, in: European Conference on Computer Vision, Springer, 2020, pp. 275–292.
[32] C. Li, R. Cong, Y. Piao, Q. Xu, C.C. Loy, RGB-D Salient object detection with cross-modality modulation and selection, in: European Conference on Computer Vision, Springer, 2020, pp. 225–241.
[33] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1724–1734.
[34] N. Ballas, L. Yao, C. Pal, A. Courville, Delving deeper into convolutional networks for learning video representations, 2015, arXiv preprint arXiv:1511.06432.
[35] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, Rgbd salient object detection: a benchmark and algorithms, in: European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[36] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: 2014 IEEE International Conference on Image Processing, ICIP, IEEE, 2014, pp. 1115–1119.
[37] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 454–461.
[38] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: Proceedings of International Conference on Internet Multimedia Computing and Service, 2014, pp. 23–27.
[39] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2075–2089.
[40] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7254–7263.
[41] N. Liu, N. Zhang, L. Shao, J. Han, Learning selective mutual attention and contrast for RGB-D saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1–16.
[42] Y. Mao, J. Zhang, Z. Wan, Y. Dai, A. Li, Y. Lv, X. Tian, D.-P. Fan, N. Barnes, Transformer transforms salient object detection and camouflaged object detection, 2021, arXiv preprint arXiv:2104.10127.

[43] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, S. Kwong, Going from RGB to RGBD saliency: A depth-guided transformation model, IEEE Trans. Cybern. 50 (8) (2019) 3627–3639.

[44] N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4722–4732.

[45] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, RGB-D Salient object detection: A survey, Comput. Vis. Media (2021) 37–69.

[46] L. Tang, Cosformer: Detecting co-salient object with transformers, 2021, arXiv preprint arXiv:2104.14729.

[47] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, C. Hou, An iterative co-saliency framework for RGBD images, IEEE Trans. Cybern. 49 (1) (2017) 233–246.

[48] G. Li, C. Zhang, G. Lin, CycleSegNet: OBject co-segmentation with cycle refinement and region correspondence, IEEE Trans. Image Process. 30 (2021) 5652–5664.

[49] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, H. Lu, Towards high-resolution salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7234–7243.

[50] P. Zhang, W. Liu, Y. Zeng, Y. Lei, H. Lu, Looking for the detail and context devils: High-resolution salient object detection, IEEE Trans. Image Process. 30 (2021) 3204–3216.

[51] Y. Zhang, L. Zhang, W. Hamidouche, O. Deforges, CMA-Net: A cascaded mutual attention network for light field salient object detection, 2021, arXiv preprint arXiv:2105.00949.

[52] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, H. Lu, Lfnet: Light field fusion network for salient object detection, IEEE Trans. Image Process. 29 (2020) 6276–6287.

[53] Z. Liu, S. Shi, Q. Duan, W. Zhang, P. Zhao, Salient object detection for RGB-D image by single stream recurrent convolution neural network, Neurocomputing 363 (2019) 46–57.

[54] X. Zhao, L. Zhang, Y. Pang, H. Lu, L. Zhang, A single stream network for robust and real-time RGB-D salient object detection, in: European Conference on Computer Vision, Springer, 2020, pp. 646–662.

[55] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, H. Du, RGB-D Salient object detection via 3D convolutional neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1063–1071.

[56] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.

[57] Z. Liu, W. Zhang, P. Zhao, A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection, Neurocomputing 387 (2020) 210–220.

[58] Z. Chen, R. Cong, Q. Xu, Q. Huang, DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 7012–7024.

[59] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, PDNet: Prior-model guided depth-enhanced network for salient object detection, in: 2019 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2019, pp. 199–204.

[60] Y. Piao, Z. Rong, M. Zhang, W. Ren, H. Lu, A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9060–9069.

[61] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, M.-M. Cheng, MobileSal: EXtremely efficient RGB-D salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1–11.

[62] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, et al., Calibrated RGB-D salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9471–9481.

[63] Q. Chen, K. Fu, Z. Liu, G. Chen, H. Du, B. Qiu, L. Shao, EF-NEt: A novel enhancement and fusion network for RGB-D saliency detection, Pattern Recognit. 112 (2021) 107740.

[64] C. Chen, J. Wei, C. Peng, H. Qin, Depth quality aware salient object detection, IEEE Trans. Image Process. 30 (2021) 2350–2363.

[65] W. Ji, J. Li, M. Zhang, Y. Piao, H. Lu, Accurate RGB-D salient object detection via collaborative learning, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, Springer, 2020, pp. 52–69.

[66] J. Zhao, Y. Zhao, J. Li, X. Chen, Is depth really necessary for salient object detection? in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1745–1754.

[67] Y. Wang, Y. Li, J.H. Elder, R. Wu, H. Lu, L. Zhang, Synergistic saliency and depth prediction for RGB-D saliency detection, in: Proceedings of the Asian Conference on Computer Vision, 2020, pp. 336–352.

[68] M. Zhang, W. Ren, Y. Piao, Z. Rong, H. Lu, Select, supplement and focus for RGB-D saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3472–3481.

[69] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[70] B. Jiang, Z. Zhou, X. Wang, J. Tang, B. Luo, CmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks, IEEE Trans. Multimed. 23 (2020) 1343–1353.

[71] X. Zhou, H. Wen, R. Shi, H. Yin, J. Zhang, C. Yan, FANet: Feature aggregation network for RGBD saliency detection, Signal Process., Image Commun. (2021) 116591.

[72] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, G. Lin, RGBD Salient object detection via disentangled cross-modal fusion, IEEE Trans. Image Process. 29 (2020) 8407–8416.

[73] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, H. Qin, Data-level recombination and lightweight fusion scheme for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2020) 458–471.

[74] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F.S. Saleh, T. Zhang, N. Barnes, UC-NEt: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8582–8591.

[75] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token ViT: Training vision transformers from scratch on ImageNet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.

[76] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, L. Shao, RGB-D Saliency detection via cascaded mutual information minimization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4338–4347.

[77] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, L. Shao, Specificity-preserving RGB-D saliency detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4681–4691.

[78] P. Sun, W. Zhang, H. Wang, S. Li, X. Li, Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1407–1417.

[79] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[80] Y. Zhai, D.-P. Fan, J. Yang, A. Borji, L. Shao, J. Han, L. Wang, Bifurcated backbone strategy for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 8727–8742.

[81] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3907–3916.

[82] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[83] J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, Adv. Neural Inf. Process. Syst. 28 (2015) 577–585.

[84] Y. Kim, Y. Jernite, D. Sontag, A. Rush, Character-aware neural language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016, pp. 2741–2749.

[85] Y. Liu, J. Li, Q. Yan, X. Yuan, C. Zhao, I. Reid, C. Cadena, 3D gated recurrent fusion for semantic scene completion, 2020, arXiv preprint arXiv:2002.07269.

[86] N. Liu, G. Gao, X. Xu, Recurrent embedded hourglass network for single image super-resolution, IEEE Access 8 (2020) 166176–166183.

[87] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, Z. Wen, Gated recurrent fusion with joint training framework for robust end-to-end speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2020) 198–209.

[88] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, Gated recurrent fusion of spatial and spectral features for multi-channel speech separation with deep embedding representations, in: Proc. Interspeech, Vol. 2020, 2020, pp. 3321–3325.

[89] A. Pahuja, A. Majumder, A. Chakraborty, R.V. Babu, Enhancing salient object segmentation through attention, in: CVPR Workshops, 2019, pp. 27–36.

[90] S. Bardhan, S. Das, S. Jacob, Visual saliency detection via convolutional gated recurrent units, in: International Conference on Neural Information Processing, Springer, 2019, pp. 162–174.

[91] S. Bardhan, Salient object detection by contextual refinement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 356–357.

[92] Z. Shi, X. Shen, H. Chen, Y. Lyu, Global semantic consistency network for image manipulation detection, IEEE Signal Process. Lett. 27 (2020) 1755–1759.

[93] W. Wang, J. Shen, M.-M. Cheng, L. Shao, An iterative and cooperative top-down and bottom-up inference network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5968–5977.

[94] B. Li, Z. Sun, L. Tang, Y. Sun, J. Shi, Detecting robust co-saliency with recurrent co-attention neural network, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019, pp. 818–825.

[95] B. Li, Z. Sun, Q. Li, Y. Wu, A. Hu, Group-wise deep object co-segmentation with co-attention recurrent neural network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8519–8528.

[96] G. Yao, X. Liu, T. Lei, Action recognition with 3D convnet-gru architecture, in: Proceedings of the 3rd International Conference on Robotics, Control and Automation, 2018, pp. 208–213.

[97] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[98] F. Du, P. Liu, W. Zhao, X. Tang, Correlation-guided attention for corner detection based visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6836–6845.
[99] J. Wei, S. Wang, Q. Huang, F^3net: Fusion, feedback and focus for salient object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12321–12328.
[100] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.
[101] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in: International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 698–704.
[102] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4548–4557.
[103] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1597–1604.
[104] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 733–740.
[105] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[106] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, M.-M. Cheng, CDNet: Complementary depth network for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 3376–3390.
[107] N. Huang, Y. Yang, D. Zhang, Q. Zhang, J. Han, Employing bilinear fusion and saliency prior information for RGB-D salient object detection, IEEE Trans. Multimed. (2021) 1–14.