Assisting RGB And Depth Salient Object Detection with Non-Convolutional Encoder: An Improvement Approach

Shuo Zhang[a], **Mengke Song**[a], **Luming Li**[a,*]
[a]College of Computer Science and Technology, China University of Petroleum (East China), Oingdao, China

Abstract. RGB-D Salient Object Detection (SOD) is a challenging task in computer vision, and deep architectures have been widely adopted in previous studies. However, current convolutional neural network (CNN)-based models struggle with capturing global long-distance features efficiently, while Transformer-based methods are computationally intensive. To address these limitations, we propose a non-convolutional feature encoder. This encoder captures long-distance dependencies while reducing computation costs, making it a potential alternative to CNNs and Transformers. Additionally, we introduce a spatial info enhancing mechanism to overcome weakened local information while capturing long-range dependencies. This mechanism balances local and global information at different expansion rates by exploring multi-scale feature fusion in the feature maps. Furthermore, we introduce a spatial info sensing module to enhance the compatibility of multi-modal features in long-range dependencies and extract informative cues from depth features. Through comprehensive experiments on four widely used datasets, we demonstrate that our proposed Involution Encoder significantly outperforms previous state-of-the-art RGB-D salient object detection methods based on CNNs in four key metrics. Compared to Transformer-based methods, our approach balances speed and efficiency favorably.

Keywords: RGB-D Salient Object Detection, non-convolution feature encoder, multi-scale feature fusion, long-distance dependency, multi-modal features..

*Corresponding author, liluming1224@126.com

1 Introduction

RGB and Depth Salient Object Detection (RGB-D SOD) is an essential and important task in computer vision, which aims to detect and highlight the most salient objects in images RGB and Depth. It is useful in many computer vision tasks, *e.g.*, object segmentation,[1–3] tracking,[4–6] image/video compression,[7–9] autonomous driving, augmented reality, and robotics. Previous works mainly rely on sole RGB images to detect salient regions, called RGB SOD,[10] which has been proven to be limited in some scenarios, such as similar foreground and background, cluttered/complex background, or low-contrast environments.

As the depth cameras develop, depth information can be a supplement to help locate salient regions more accurately. Most recent deep learning-based fusion methods can be categorized into three types: 1) input fusion, 2) late fusion, and 3) mid fusion. Though input and late fusion have advantages, they usually perform very poorly due to the absence of feature interaction. Thus the current mainstream SOTA models[11,12] have concentrated more on mid fusion to mine how to integrate RGB cues and depth (D) cues more sufficiently and completely.

Nevertheless, merely concerning the fusion process maybe not be enough, which has presumably overlooked the global context information when extracting features. Because current typical encoders, such as ResNet[13] and VGG,[14] are based on the CNN architecture, which is weak in modeling long-distance dependencies and capturing the large receptive fields. Also, the information between channels is redundant. As DETR[15] introduces Transformer from Natural Language Processing to Computer Vision, Transformer-based encoders become increasingly popular. It's a non-local model with self-attention and cross-attention layer to capture long-range dependencies in an image and has helped Transformer-based RGB-D methods achieve excellent results. In addressing the computational resource requirements of Transformer-based methods and their impact on efficiency and practicality, recent efforts have explored alternative architectures. For instance, Li et al.[16] proposed Involution, a non-convolutional architecture that utilizes involution kernels generated based on individual pixels rather than connections with neighboring pixels. Even relatively simple involution structures can achieve a competitive balance between accuracy and computational cost.

Motivated by Involution, this paper introduces a novel and efficient non-convolutional feature encoder network (NCFE-Net). NCFE-Net stands out due to its unique design, which integrates involution into existing convolution-based feature encoders, transforming them into non-convolutional feature encoders. This design enables the model to capture long-range dependencies with minimal computational requirements. To further address the issue of weakened local information during the capture of long-range dependencies, NCFE-Net incorporates a spatial info enhancing mechanism (SIE). By automatically exploring multi-scale information in feature maps, SIE balances incorporating local and global information at different expansion rates, leading to improved model performance. Additionally, to enhance the compatibility of multi-modal features and improve the expressive power of depth features in capturing long-range dependencies, NCFE-Net integrates a spatial info sensing module (SIS). This module refines and strengthens the input multi-modal features, extracting more informative clues and effectively enhancing the model's overall performance.

To summarize, the main contributions of this work are four-fold:

- A novel non-convolutional feature encoder is designed to capture long-range dependencies while reducing computational requirements, achieving a balance between speed and accuracy;
- A novel and effective spatial info enhancing mechanism is proposed, which explores multi-scale feature fusion and ensures a balance between local and global information at different sampling rates within the feature maps;
- A spatial info sensing module is introduced to enhance the compatibility of multi-modal features and extract informative clues from depth features in capturing long-range dependencies more effectively;
- Extensive experiments are conducted on four publicly available datasets, demonstrating the effectiveness and superior performance of the proposed network; Both codes and results will be publicly available, which has the potential to benefit our research community in the near future.

2 Related Work

2.1 CNN-based RGB-D salient object detection

Traditional methods in image saliency detection heavily rely on handcrafted features[17–32] and incorporate various saliency priors, such as contrast priors, image background priors, and object priors. In 2017, Zhu *et al.*[18] utilized the center-dark channel prior method, which generates a center-dark channel mapping by computing center saliency priors and dark channel priors. The initial saliency map is then fused with the center-dark channel mapping to obtain the final saliency map. In 2018, Zhu *et al.*[17] introduced a deep mining-based multi-layer backpropagation saliency detection algorithm that utilizes depth cues from three different levels of the image. However, these methods overlook the inherent differences between RGB and depth modalities, leading to potentially unreliable results, particularly in detecting small objects.

The advent of deep learning has revolutionized the field, with convolutional neural network (CNN) based methods[33–40] taking the lead. Among them, fusion methods [41–45] have made significant strides in RGB-D saliency detection and achieved remarkable performance. Notably, in 2020, Li *et al.*[50] proposed an interactive adaptive fusion method that enhances high-level RGB and depth features, distinguishing cross-modal features from different sources and reinforcing RGB features with depth features at each level. Cong *et al.*[48] introduced a metric to assess their reliability and utilized it for merging two prediction results. Song *et al.*[46] performed multi-scale pre-segmentation on RGB-D pairs and proposed a multi-scale discriminative saliency fusion method to generate the final saliency map. For late fusion, Guo *et al.*[47] iteratively propagated the initial saliency map obtained through multiplication to produce the final saliency map. To account for the quality of depth maps, Moreover, for mid-level fusion, Fan *et al.*[49] employed a dual-stream structure to transform cross-modal features and fuse cross-layer features, explicitly filtering out low-quality depth maps using a gating mechanism. In 2021, Chen *et al.*[51] integrated a depth quality perception sub-network into a classical dual-stream structure and assigned weights to depth features before fusion, facilitating effective RGB and depth information fusion.

However, most current saliency detection methods are primarily based on CNN architectures, which limit their ability to capture long-range dependencies. Some methods integrate global and local information to achieve accurate salient region detection. For example, Zhang *et al.*[52] proposed a framework that considers the complementarity of global positions and local details from two modalities, yielding good results. However, these methods still struggle to fully capture the advantageous relationships between features. To address these limitations, a novel feature encoder is proposed, which utilizes a non-convolutional encoder to capture global context and efficiently performs multi-scale feature fusion using an effective spatial info enhancing mechanism within the feature maps.

2.2 Transformer-based RGB-D salient object detection

The transformer was first proposed by.[53] Once proposed, it quickly occupies a dominant position in Natural Language Processing (NLP), which is used to model global long-range dependencies, constantly refreshing records one after another. Building upon its success in various domains, including natural language processing, the Transformer architecture has recently been extended into computer vision, yielding remarkable results and solidifying its position. A crucial component within the Transformer architecture is self-attention, which plays a pivotal role in capturing robust features with long-range information by leveraging the interaction between feature self-information and weighted matrices. For instance, in 2020, Liu *et al.*[54] proposes a hierarchical

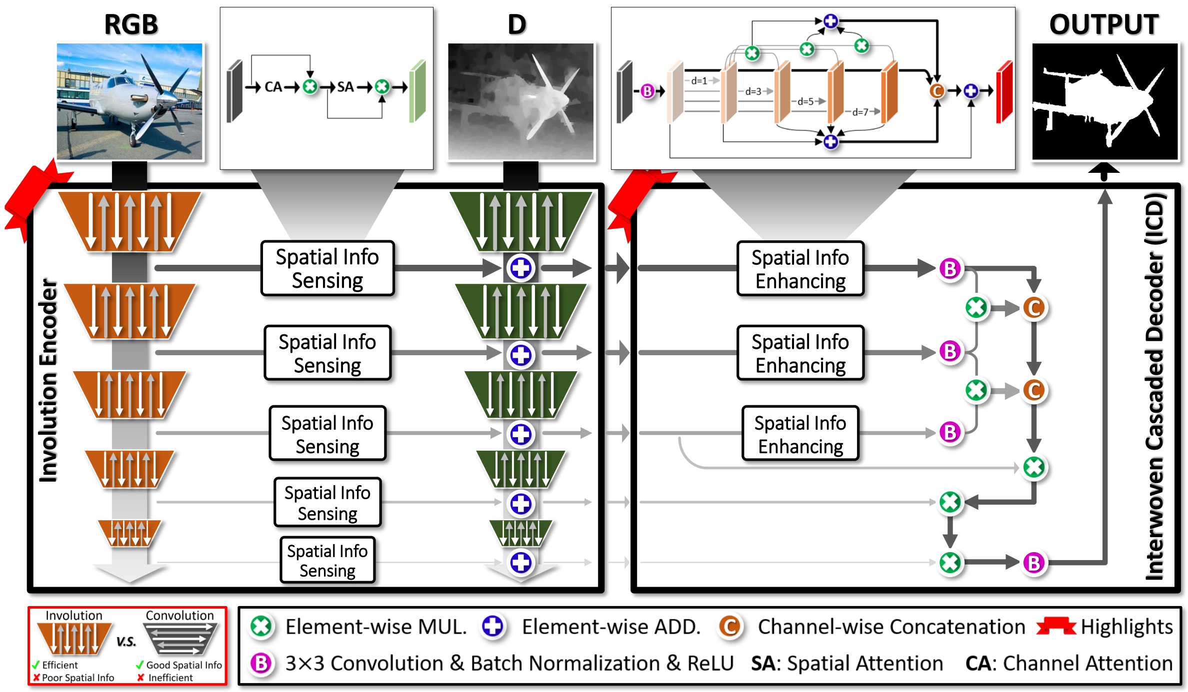


Fig 1 Method pipeline of our approach. The major highlight of our approach is the proposed non-convolution feature encoder, *i.e.*, involution encoder, to solve the limitations of standard CNN in modeling long-distance dependencies and capturing the large receptive fields.

transformer with a shift window scheme. In 2021, Liu *et al.*[55] propose to make the model transmit more effectively across window resolution. Liu *et al.* [56] propose the triple transformer embedded module to learn cross-layer long-range dependencies to enhance high-level features. Tang *et al.*[57] propose to capture significant and common visual patterns from multiple images. Ren *et al.* [58] propose a pure transformer-based encoder and a hybrid decoder to aggregate the features generated by the transformer. In 2022, Wang *et al.*[59] introduced a Transformer-based network to address the challenges of local operations in multi-scale and multi-modal fusion and capturing long-range dependencies. Although these methods have achieved performance improvements, they come at a significant computational cost. Some methods combine CNN and Transformer but may still encounter computational challenges. In contrast, the proposed novel feature encoder maintains competitive detection results while reducing computational costs.

3 Proposed Method

3.1 Overview

As is shown in Fig. 1, the key idea of our NCFE-Net is to replace the CNN-based encoder with a non-convolution feature encoder to make up for the limitations of the standard CNN in modeling long-distance dependencies and capturing the large receptive fields, which includes three main components: 1) dual-stream involution encoder (InEn); 2) spatial info enhancing (SIE); 3) spatial info sensing module (SIS). Details can be seen in the following Sec. 3.2, Sec. 3.4 and Sec. 3.3.

3.2 Involution Encoder

Existing backbones of RGB-D SOD methods mainly consist of encoder-decoder architectures, which are dominated by CNN-based networks, *e.g.*, ResNet[13] and VGG.[14] Nevertheless, as shown in Fig. 1 (left bottom), CNN has there major limitations: 1) fixed convolutional kernel sizes and strides, which may result in information loss or redundancy in certain scenarios where the receptive field is not flexible enough; 2) they can capture better local features but struggle with global features in images, failing to capture long-range spatial dependencies; 3) they are highly sensitive to position and cannot capture rich feature representations on different orientations and scales, making them vulnerable to distortions caused by image rotations and flips, leading to distorted feature representations. These inherent limitations of CNNs have resulted in insufficient global context modeling and feature representation capabilities in most existing methods. Additionally, while Transformer-based RGB-D methods have achieved excellent results in capturing long-range spatial dependencies and rich feature representations, their main drawback is the requirement of substantial computational resources, limiting their practical efficiency.

To address these limitations, a novel non-convolutional feature encoder called Involuton Encoder (InEn) is proposed, primarily utilizing involution kernel.[16] Compared to CNN, involution (Fig. 1 (left bottom)) can capture crucial features on local receptive fields of different sizes and orientations, enabling the learning of more abstract and complex feature representations, thereby enhancing the feature representation capability. Moreover, involution can adaptively adjust the convolutional kernel sizes and strides to accommodate various receptive field control requirements, effectively handling global image features.

Specifically, the output feature map of involution is derived by performing multiply-add operations on the input with involution kernels, which can be defined as:

$$\mathbf{Y}_{i,j,k} = \sum_{(u,v)\in\triangle k} \mathbf{H}_{i,j,u+\lfloor k/2\rfloor,v+\lfloor k/2\rfloor,\lceil kG/C\rceil}\mathbf{X}_{i+u,j+v,k}, \tag{1}$$

where $\mathbf{X}$ denotes the input feature map, and $\mathbf{Y}$ is the output feature map. $\triangle k$ refers to the set of offsets in the neighborhood considering convolution conducted on the center pixel, and $\mathbf{H}$ represents involution kernels. Unlike convolution kernels, the shape of involution kernels H depends on the input feature map $\mathbf{X}$.

To be more precise, the computation process of the involution operation can be divided into two main steps: generating the involution kernel and performing the involution convolution.

1) Generating the involution kernel: During the involution kernel generation step, all channel pixels at a particular spatial position are selected. These selected pixels undergo a transformation function and are then unfolded to obtain the involution kernel. This process ensures the creation of an effective involution kernel for subsequent convolution operations.

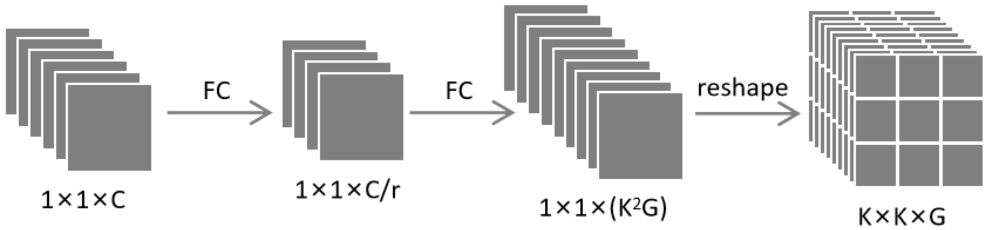


Fig 2 Visualization of the involution kernel generation process.

$$\begin{aligned} \mathbf{X}_{i,j} : 1 \times 1 \times C \overset{\textbf{FC}}{\rightarrow} 1 \times 1 \times C/r \overset{\textbf{FC}}{\rightarrow} \\ 1 \times 1 \times \left(K^2 G\right) \overset{\textbf{reshape}}{\rightarrow} \mathbf{H} : K \times K \times G \end{aligned} \tag{2}$$

where **FC** represents fully connected operation, and **reshape** denotes reshaping operation. The symbols C, G, K, and r represent the number of channels, the number of groups, the kernel size, and the scaling factor, respectively. For a better understanding of the involution kernel generation. We have included a generation process diagram in the revised manuscript in Fig. 2.

2) Computing involution: a multiply-add operation is performed, *i.e.*, the involution kernel is firstly reshaped into a matrix, and then it is element-wise multiplied with the corresponding positions of the input feature map. Finally, all the $K \times K$ elements of each channel are summed to replace the original pixel at that position.

To construct the entire network using involution, we borrowed the design from ResNet and implemented it by stacking residual blocks. In the backbone of ResNet, the convolutions are replaced with involutions, while retaining all the convolutions for channel projection and fusion. These carefully redesigned components together form the non-convolutional backbone network, which is referred to as the Non-Convolution Feature Encoder Network (NCFE-Net). Then, we take the place of the convolution kernels with involution kernels in our encoders, *e.g.*, ResNet50, and build an involution-based feature encoder.

3.3 Spatial Info Sensing

There are two main problems when trying to fuse RGB and depth features. One is the compatibility of the two due to the intrinsic modality difference, and the other is the redundancy and noise in low-quality depth features. To address these issues, spatial info sensing (SIS) is proposed. The SIS module aims to enhance the compatibility of multimodal features in capturing long-range dependencies and extracting informative cues from the depth features.

Specifically, as shown in Fig. 1, SIS consists of a channel attention $\mathbf{CA}(\cdot)$ and a spatial attention $\mathbf{SA}(\cdot)$, which captures long-range dependencies and extracting informative cues from the depth features in channel dimension and spatial dimension, being defined as:

$$F_{DR}(f_i) = \big(f_i \otimes \mathbf{CA}(f_i)\big) \otimes \mathbf{SA}\big(f_i \otimes \mathbf{CA}(f_i)\big), \tag{3}$$

where f_i denotes the ith output feature of the depth encoder.

The channel attention captures long-range dependencies for multimodal features by channel selection, which is achieved by the channel dimension's global max pooling operation ($\mathbf{GMP}_c(\cdot)$), a multi-layer perception (MLP($\cdot$)), and channel-wise multiplication (moc($\cdot$)). For input features f, global max pooling operation retains their key channel information, then multi-layer perception selects the important channel information among them. Finally, channel-wise multiplication are performed to select import channel. The channel attention $\mathbf{CA}(\cdot)$ can be represented as:

$$\mathbf{CA}(f) = moc\Big(f, \mathbf{MLP}\Big[\mathbf{GMP}_c(f)\Big]\Big), \tag{4}$$

where f denotes the input feature, $\mathbf{GMP}_c$ is the global max pooling operation over the input feature slice, MLP stands for a multi-layer perception, and $moc(\cdot,\cdot)$ performs channel-wise multiplication between its input.

The spatial attention $\mathbf{SA}(\cdot)$ enhances the compatibility of multimodal features in extracting informative cues from the depth features, which is achieved by pixel-wise global max pooling operation ($\mathbf{GMP}_s$ $(\cdot)$), the convolution operation ($\mathbf{Conv3}(\cdot)$), and element-wise multiplication ($ewm(\cdot,\cdot)$). For input features, pixel-wise global max-pooling operation down-samples them for reducing compute cost, the convolution operation extracts their spatial information. Finally, the element-wise multiplication obtains import spatial information clues. The spatial attention $\mathbf{SA}(\cdot)$ can be represented as:

$$\mathbf{SA}(f) = ewm\Big(f, \mathbf{Conv3}\Big[\mathbf{GMP}_s(f)\Big]\Big), \tag{5}$$

$\mathbf{GMP}_s$ is the pixel-wise global max-pooling over the entire input feature tensor, $\mathbf{Conv3}$ is a 3×3 convolution, and $ewm(\cdot,\cdot)$ performs element-wise multiplication between inputs.

3.4 Spatial Info Enhancing

The non-convolutional feature encoder, *i.e.*, InEn, can capture long-range dependencies in features by using the involution operation to model global information. However, this process weakens the local information of the features. A potential solution to this issue is using Atrous Spatial Pyramid Pooling (ASPP), which captures multi-scale contextual information by employing dilated convolutions with different expansion rates. Nevertheless, ASPP fails to fully exploit the relationship between features with different expansion rates by simply concatenating features at all dilation rates.

To overcome this limitation and improve the capture of intrinsic relationships between features by balancing local and global information at different expansion rates, a new method named spatial info enhancing (SIE) is introduced. The SIE aims to fully utilize the multi-scale feature fusion of contextual information in the feature map by considering the relationship between features with different expansion rates. By incorporating SIE, the model can better capture and balance the intrinsic relationships between features using different dilation rates while optimizing the available information in the architecture.

The traditional ASPP method involves five parallel branches. Initially, a 1×1 convolution is applied to all branches, which reduces the channel size to 32. The first branch then performs two consecutive convolution operations with kernel sizes of 3 and expansion rates of 1, 3, 5, and 7, respectively. The whole process can be denoted as follows:

$$\mathbf{ASPP}(f) = \mathbf{Concat}\big(\underset{\overbrace{\mathbf{Conv1}(f)}}{\underset{\Uparrow}{f^{'}}}, \mathbf{DConv3}_{d=i}(\underset{\overbrace{\mathbf{Conv1}(f)}}{\underset{\Uparrow}{f^{'}}})\big), \tag{6}$$

where $\mathbf{DConv3}_{d=i}(\cdot)$ is the dilated convolution with expansion rates $d = i$ $(i \in 1, 3, 5, 7)$. f is the input feature. To implement SIE, two branches are utilized, as depicted in Fig. 1. The first branch involves element-wise multiplication between features with dilation rates of 3, 5, and 7 and features expanded at a rate of 1. The output features are then added together, which is defined as:

$$f_1^{'} = \sum_{i \in \{3,5,7\}} \mathbf{DConv3}_{d=1}(f^{'}) \otimes \mathbf{DConv3}_{d=i}(f^{'}), \tag{7}$$

where $\sum$ denotes the element-wise summation and $\otimes$ is the element-wise multiplication. The second branch directly adds up the four features expanded at different rates. Finally, the features from both branches are concatenated to generate the final output feature map $f_r^{'}$, which can be defined as:

$$f_2^{'} = \sum_{i \in \{1,3,5,7\}} \mathbf{DConv3}_{d=i}(f^{'}), \tag{8}$$

$$f_r^{'} = \sum \Big(f, \mathbf{Concat}\big(f_1^{'}, f_2^{'}, \mathbf{Conv1}(f)\big)\Big). \tag{9}$$

Notice that our SIS can jointly excavate informative cues from depth features in multiple side-out layers. Component experiments (see Table 2) show the effectiveness of this approach in improving the compatibility of multi-modal features.

The overall operation flow reveals that features with a dilation rate of 1 are used multiple times compared to features with expansion rates of 3, 5, and 7. This is because features with an expansion rate of 1 focus more on local information, whereas those with expansion rates of 3, 5, and 7 are sparser and concentrate more on global information. To balance the fusion of information between global and local scales, this spatial info enhancing fusion method balances multi-scale feature fusion and contextual information within the feature map. This significantly enhances the performance of ASPP and addresses the issue of weakened local information of features while capturing long-range dependencies.

3.5 Interwoven Cascaded Decoder

The multilevel cross-modal features computed from NCFE-Net are a fusion of RGB and depth features from multiple levels. To effectively utilize the multi-scale and multilevel information within each level for cascaded refinement, a lightweight decoding mechanism called the interwoven cascaded decoder (ICD) has been implemented to integrate the multilevel cross-modal features. As illustrated in Fig. 1, the ICD comprises three spatial info enhancing modules and a straightforward feature aggregation strategy consisting of cascaded convolutions, element-wise multiplications and channel-wise concatenations to extract global contextual information from cross-modal features.

Compared to existing decoders, the ICD can simultaneously process multiple levels of information by utilizing multilevel information from both RGB and depth modalities. This allows the model to capture spatial and contextual information more effectively, leading to more accurate saliency predictions. The ICD consists of multiple stages, each responsible for aggregating information from different levels and modalities. Furthermore, the decoder has a cascading structure that enables the features from the previous layer to serve as inputs for subsequent stages. As information propagates through the decoder, predictions are iteratively refined, improving accuracy. In addition to its cascading structure, the ICD introduces an interweaving mechanism that helps to better fuse information from RGB and depth modalities. This mechanism leverages the differences in modality characteristics, allowing the model to capture complementary information better. In essence, the ICD decoder is a highly effective tool for improving the performance of RGB-D saliency detection models because it can process multilevel information and interweave information from different modalities. This results in better feature fusion and more accurate saliency prediction, making it a valuable asset to researchers and practitioners.

4 Experiments

4.1 Datasets

We evaluate the effectiveness of our model on four widely used public benchmark datasets, *i.e.*, NJUD,[60] NLPR,[61] SIP,[49] STEREO.[62] NJUD[60] includes 2,003 stereo image pairs with various resolutions. Among these image pairs, 1,400 are used as the training set, 100 as the validation set, and the remaining as the testing set. NLPR[61] consists of 1,000 images from 11 indoor and outdoor scene types. Among them, 650 images are used as the training set, 50 images as the validation set, and the remaining 300 images as the testing set. SIP[49] consists of 1,000 high-resolution images that cover diverse real-world scenes from various viewpoints, poses, occlusions, illuminations, and backgrounds. STEREO[62] has 797 stereoscopic images. These images are mainly collected from the Internet and 3D movies. Depth images are generated by leveraging an optical method. Evaluating the proposed model on these datasets can validate its effectiveness, and its performance can be compared and analyzed objectively.

4.2 Evaluation Metrics

Three metrics are adopted for quantitative evaluation, including **S**-measure (Sm),[63] **F**-measure (Fm),[64] and mean absolute error (MAE). Specifically, S-measure is utilized to solve the problem of structural measurement from the perspective of region-aware and object-aware. F-measure offers a unified solution to evaluating non-binary and binary maps. The MAE denotes the average pixel-wise difference between saliency maps and the ground truth. These metrics can comprehensively evaluate the model's performance in the saliency detection task. F-measure is an important performance indicator when precision rate conflict with recall rate, and it can be computed as Eq. 10, which shows the balance between precision rate and recall rate:

$$Fm = \frac{(\beta^2 + 1) \times PRE \times REC}{\beta^2 \times PRE + REC}, \tag{10}$$

where PRE represents the average precision rate, REC represents the average recall rate, and $\beta^2 = 0.3$ to balance the precision rate and the recall rate. S-measure is also called Structure-measure. The novel evaluation focuses on the region-wise and object-wise structural similarities, which is

Table 1 Quantitative comparison with current SOTA models on four widely-used datasets in terms of S, F_β and MAE (M). ↑ means that the larger the numerical value, the better the model, while ↓ means the opposite. The best results are marked in **bold**.

Datasets			NJUD			NLPR			SIP			STEREO		
Methods			S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓
CNN-based	PCA	2018	0.877	0.844	0.059	0.873	0.794	0.044	0.842	0.824	0.071	0.880	0.845	0.061
	CPFP	2019	0.878	0.877	0.053	0.888	0.822	0.036	0.850	0.818	0.061	0.871	0.827	0.054
	DMRA	2019	0.886	0.872	0.051	0.899	0.855	0.031	0.806	0.819	0.085	0.886	0.868	0.047
	cmMS	2020	0.900	0.897	0.044	0.915	0.896	0.027	0.872	0.877	0.058	0.895	0.879	0.043
	ICNet	2020	0.894	0.843	0.052	0.923	0.908	0.028	0.854	0.791	0.070	0.891	0.847	0.046
	SSF	2020	0.899	0.896	0.043	0.914	0.896	0.026	0.878	0.880	0.054	0.887	0.882	0.046
	ATSA	2020	0.901	0.893	0.040	0.907	0.876	0.028	0.864	0.873	0.058	0.897	0.884	0.039
	UCNet	2020	0.897	0.895	0.043	0.92	0.901	0.025	0.875	0.876	0.051	0.903	**0.899**	0.039
	BBSNet	2021	0.919	0.899	0.037	0.926	0.878	0.028	0.874	0.874	0.056	0.909	0.886	0.041
	ASIF	2021	0.889	0.888	0.047	0.906	0.888	0.030	0.857	0.859	0.061	0.868	0.893	0.049
	MAD	2022	0.921	0.903	0.037	0.933	0.901	0.026	0.884	0.877	0.051	0.910	0.892	**0.037**
	Mobilesal	2022	0.905	**0.914**	0.041	0.920	0.907	0.025	0.873	**0.882**	0.053	0.895	0.891	0.045
Ours			**0.925**	0.905	**0.033**	**0.930**	**0.910**	**0.024**	**0.891**	**0.882**	**0.049**	**0.911**	**0.899**	**0.037**
Transformer-based	GROUPTrans	2022	0.922	**0.921**	0.028	0.928	0.908	0.019	0.887	**0.895**	**0.041**	0.908	0.895	0.032
	CAVER	2023	0.920	0.900	0.031	0.929	0.895	0.022	**0.893**	0.868	0.042	**0.914**	0.883	0.033
Ours			**0.925**	0.905	**0.033**	**0.930**	**0.910**	0.024	0.891	0.882	0.049	0.911	**0.899**	0.037

more similar to the human visual system. It can be formulated as:

$$Sm = \alpha \times S_o + (1 - \alpha) \times S_r, \tag{11}$$

where we set $\alpha = 0.5$ to balance the region-aware (Sr) and object-aware (So) structural similarity. The MAE is defined as:

$$\text{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - GT(x,y)|, \tag{12}$$

where W and H respectively represent the image width and image height; P represents the estimated saliency map and GT denotes the ground truth.

4.3 Comparison with state-of-the-art models

To demonstrate the effectiveness of the proposed method, we compare it with 12 state-of-the-art (SOTA) CNN-based RGB-D SOD methods, *i.e.*, PCA,[47] CPFP,[65] DMRA,[66] ICNet,[48] SSF,[52] ATSA,[67] UCNet,[12] BBSNet,[68] ASIF,[69] MAD,[70] MobileSal,[71] and two Transformer-based RGB-

D SOD methods, *i.e.*, GroupTransNet,[72] CAVER.[73] The compared results are from the codes or saliency maps provided by the authors. The quantitative comparison results are shown in Table 1. It can be seen that our method performs the best on NJUD and NLPR datasets and shows competitive performance on STEREO and SIP datasets, which proves the effectiveness of the proposed NCFE-Net model. In particular, in terms of the S metric, our method consistently outperforms all other compared SOTA methods, *e.g.*, 0.891 (ours) v.s. 0.878 in the SIP set. The superiority of our RGB-D Salient Object Detection method in terms of the S metric stems from the unique design of our non-convolutional feature encoder, which efficiently captures long-distance dependencies. Unlike CNN-based models that struggle with global feature representation, and Transformer methods that are computationally heavy, our encoder is optimized for both efficiency and effectiveness. Additionally, our spatial info enhancing mechanism adeptly balances local and global information, utilizing multi-scale feature fusion for a more refined saliency detection. The spatial info sensing module further augments this by ensuring multi-modal features harmonize over long ranges and by extracting salient cues from depth features more effectively than existing methods. These innovations collectively contribute to our method's exceptional performance on standard benchmarks. Also, we can find that our method outperforms all compared CNN-based RGB-D SOD methods. Our method shows competitive results on Transformer-based RGB-D SOD methods and achieves the trade-off between speed and efficiency at the same time, which would be discussed in Section 4.5.3.

Fig. 3 presents visual comparison results of NCFE-Net with state-of-the-art representative models, highlighting the excellent performance of the proposed model in detecting single objects in low-contrast images in the first row. The second, third, and fourth rows show that the proposed model outperforms others in capturing salient regions with more complex objects, resulting in

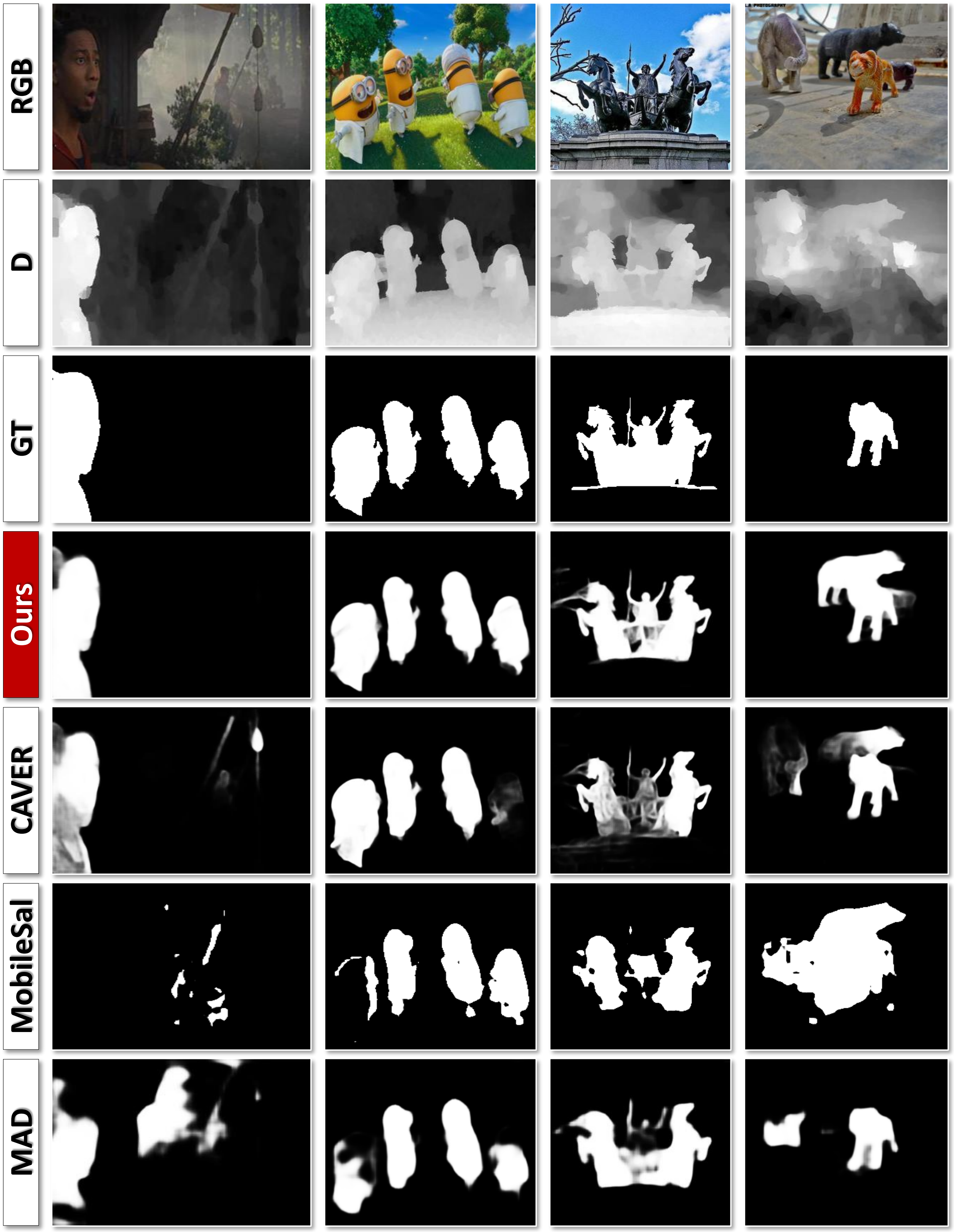


Fig 3 Visual comparison between our method and several most representative SOTA models.

clear boundaries. These results demonstrate the effectiveness of NCFE-Net in saliency detection, particularly in scenarios involving complex backgrounds and objects of different shapes.

4.4 Component Evaluation

We conducted an extensive component evaluation to confirm the major components' effectiveness in our approach, as shown in Table 2. The results indicate that all components of our proposed algorithm contribute to improving the saliency detection performance.

To provide more specific details, InEn (Sec. 3.2) plays a crucial role in capturing long-range dependencies and reducing model computation costs. It offers a viable alternative to CNN and Transformer architectures, and empirical evidence demonstrates its significant impact on improving saliency detection performance. For instance, on the NJUD dataset, adopting NCFE increased the S metric from 0.869 to 0.899.

Furthermore, including SIE (Sec. 3.4) has positively influenced the model's performance. By incorporating multi-scale feature fusion in feature maps and balancing local and global information at different expansion rates, SIE effectively handles features with varying sampling rates, improving prediction accuracy. Experimental results reveal that replacing ASPP with SIE further

Table 2 Components evaluation of S, F_β and MAE(M) on the NJUD and NLPR dataset. The best results are marked in **bold**. Where, Ba denotes baseline (CNN encoder). InEn denotes involution encoder. SIS denotes spatial info sensing. SIE denotes spatial info enhancing. ASPP denotes atrous spatial pyramid pooling. R denotes our final version.

	Key Components					Datasets					
	Ba	InEn	SIS	SIE	ASPP	NJUD			NLPR		
						S↑	F_β↑	M↓	S↑	F_β↑	M↓
1	✓	✗	✗	✗	✓	0.869	0.804	0.084	0.881	0.857	0.076
2	✗	✓	✗	✗	✓	0.899	0.873	0.061	0.892	0.877	0.065
3	✗	✓	✗	✓	✗	0.910	0.890	0.061	0.915	0.894	0.039
4	✗	✓	✓	✗	✓	0.905	0.887	0.050	0.901	0.885	0.041
5	✓	✗	✓	✓	✗	0.916	0.892	0.043	0.919	0.902	0.036
R	✗	✓	✓	✓	✗	**0.925**	**0.905**	**0.033**	**0.930**	**0.910**	**0.024**

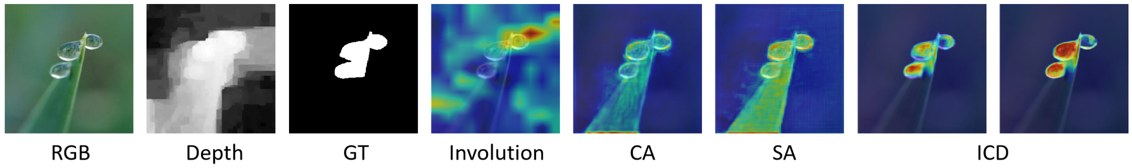


Fig 4 Visualization of the proposed components.

enhances the S metric on the NJUD dataset by 1.1%.

While SIS (Sec. 3.3) marginally enhances performance, its contribution is smaller than SIE. For example, when applied to the NLPR dataset, using SIS improved the S-measure from 0.892 to 0.901, whereas using SIE increased the S metric to 0.915, highlighting the superiority of SIE in enhancing cross-receptive spatial feature fusion.

Finally, integrating the non-convolutional encoder with ASPP and SIE can significantly boost the overall model performance. Comparative experimental results demonstrate that the non-convolutional encoder works more effectively within the framework, underscoring its superior performance as a key component.

In summary, the results of the component evaluation confirm that all proposed components significantly contribute to enhancing saliency detection performance. These findings highlight the importance of carefully selecting appropriate components and recognizing their impact on the overall algorithm's performance. Notably, ASPP, SIE, and the involution encoder are key components that play a crucial role in improving performance. These findings emphasize the significance of selecting suitable components to develop high-performance saliency detection models.

4.5 Ablation study

4.5.1 Different Fusion Methods of SIE

To assess the effectiveness of the proposed spatial info enhancing (SIE, Sec. 3.4) method, three experiments were conducted. These experiments compared only element-wise addition with different expansion rates ("Replace MUL."), only element-wise multiplication ("Replace ADD."), or a combination of the two while keeping the parameters consistent. The performance of each feature fusion operation was recorded and presented in Table 3, which indicates that the combination of both operations (the proposed SIE method) produced the best performance. By contrast, the original ASPP method performed the worst, illustrating the efficiency and effectiveness of the proposed SIE method. This finding is reasonable because fusing feature maps with varying expansion rates can provide complementary information, thereby enhancing the representative ability of the model's features.

Table 3 Ablation results of different fusion methods compared with spatial info enhancing (SIE). The best results are marked in **bold**. "Replace ADD." denotes replace all addition operations by multiplication operations; "Replace MUL." denotes replace all multiplication operations by addition operations.

Datasets	NJUD			NLPR			SIP			STEREO		
Choices	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓
Classic ASPP	0.905	0.887	0.050	0.901	0.885	0.041	0.876	0.860	0.058	0.895	0.871	0.047
Replace MUL.	0.919	0.897	0.038	0.922	0.896	0.030	0.885	0.878	0.053	0.898	0.882	0.045
Replace ADD.	0.921	0.901	0.035	0.924	0.904	0.027	0.889	0.878	0.052	0.904	0.887	0.042
The Proposed SIE	**0.925**	**0.905**	**0.033**	**0.930**	**0.910**	**0.024**	**0.891**	**0.882**	**0.049**	**0.911**	**0.899**	**0.037**

4.5.2 Ablation study on Interwoven Cascaded Decoder

To further evaluate the effectiveness of the proposed interwoven cascaded decoder (ICD, Sec. 3.5), we conducted additional experiments by comparing it with two alternative decoding mechanisms: element-wise addition and element-wise multiplication. In the element-wise addition mechanism, only element-wise addition is used to fuse features from different layers. In contrast, only element-wise multiplication is employed in the element-wise multiplication mechanism.

Table 4 Effectiveness analysis of the interwoven cascaded decoder (ICD). The best results are marked in **bold**. "All ADD./MUL. Operations" means all fusion operations in ICD are replaced by Addition/Multiplication operation.

Datasets	NJUD			NLPR			SIP			STEREO		
Choices	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓	S ↑	F_β ↑	M ↓
All ADD. Operations	0.915	0.895	0.047	0.921	0.902	0.028	0.888	0.865	0.058	0.901	0.879	0.048
All MUL. Operations	0.911	0.889	0.042	0.918	0.897	0.032	0.885	0.870	0.053	0.903	0.886	0.043
The Proposed ICD	**0.925**	**0.905**	**0.033**	**0.930**	**0.910**	**0.024**	**0.891**	**0.882**	**0.049**	**0.911**	**0.899**	**0.037**

In these experiments, we employed 1×1 convolutions and upsampling operations to ensure that features from different layers have the same dimensions. Subsequently, the features were fused using either element-wise multiplication ("All MUL. Operations") or element-wise addition ("All ADD. Operations"). The results of these experiments, as presented in Table 4, clearly indicate the superiority of the interwoven cascaded decoder (ICD).

The results demonstrate that the interwoven cascaded decoder outperforms both alternative decoding methods in terms of performance. This finding confirms the effectiveness of the interwoven cascaded decoder in integrating features from different layers and enhancing overall performance.

4.5.3 Comparison with Transformer-based Methods

It is worth mentioning that Transformer-based methods have demonstrated superior performance compared to the proposed NCFE-Net (Ours). However, these methods often require substantial computational resources and present challenges in terms of training. In contrast, while NCFE-Net may exhibit slightly lower performance than Transformer-based methods, it offers a favorable balance between performance, inference speed, and model size, as shown in Table 5.

Currently, Transformer-based methods dominate the field of salient object detection. However, their computational requirements make it challenging to apply them in real-time applications or on devices with limited computing power. In contrast, the proposed NCFE-Net presents a viable alternative that delivers good performance while maintaining a relatively small model size and

Table 5 Model size and speed analysis of Transformer-based methods and our non-convolutional-based method. The bests are marked in **bold**.

Competitors	UCNet	cmMS	MAD	**Ours**	GroupTrans	CAVER
Model Size	119 MB	270 MB	310 MB	**78 MB**	140 MB	115 MB
FPS	42	15	52	**51**	37	28
	CNN-based				Transformer-based	

high inference speed (Frames Per Second, FPS). This advantage is especially valuable for practical applications prioritizing high speed and efficiency.

By highlighting these considerations, it becomes evident that NCFE-Net offers a practical solution that balances performance and computational requirements, making it well-suited for real-time applications and resource-constrained environments.

Further, we have provided visual comparison between our method and other CNN-based and Transformer-based methods in terms of three challenging situations — "similar foreground and background, "cluttered/complex background", and "low-contrast environments". Results in Fig. 5 demonstrate that our method outperforms the other methods in these challenging situations.

In the case of "similar foreground and background" (line 1), our method successfully distinguishes the foreground object from the background, while the other methods struggle due to the lack of clear visual separation.

Regarding "cluttered/complex background" (line 2), our method shows superior performance by accurately detecting the object of interest amidst the complex surroundings. On the other hand, the other methods fail to achieve the same level of precision and tend to produce false positives or miss detections.

In the case of "low-contrast environments" (line 3), our method exhibits robustness by effectively detecting objects even in situations with low contrast between the object and the background. Conversely, the other methods face difficulties in detecting objects under such conditions.

4.6 Failure Cases

We present some failure cases in Fig. 6. Despite our method's promising results, two major challenges still need to be addressed. Firstly, the method struggles to extract cross-modality features

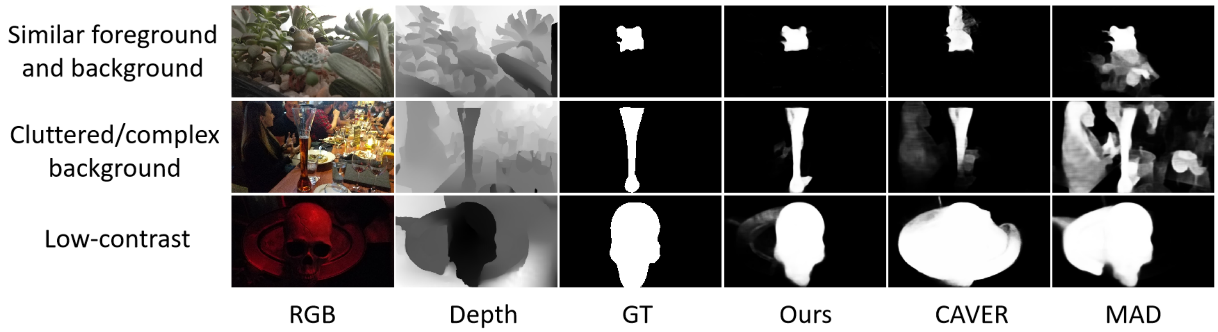


Fig 5 Visual comparison between our method and other CNN-based (MAD) and Transformer-based (CAVER) methods in terms of three challenging situations.

fully and can be easily affected by poor features from a modality. Secondly, when dealing with images with complicated backgrounds, our method may highlight only certain parts of the scene rather than the entire salient region. In situations where false-alarm salient objects are present in the depth map, such as in the first row of Fig. 6, our method may struggle to detect these objects accurately. This is due to the difficulty in fully extracting cross-modality features, which can lead to the mistaken identification of false-alarm salient objects. In the bottom row of Fig. 6, we demonstrate how our method may only highlight certain parts of a scene with a complicated background. This occurs because the method can struggle to identify the entire salient region of an image with a complex background.

While our method has shown promising results, further improvements are needed to address these challenges and improve its accuracy in difficult scenarios.

5 Conclusions

This paper introduces an innovative and effective method called the non-convolutional feature encoder network (NCFE-Net) for RGB-D salient object detection. The network leverages involution to capture long-range dependencies while maintaining a smaller computational cost than Transformers. Additionally, the approach incorporates spatial info enhancing for multi-scale feature

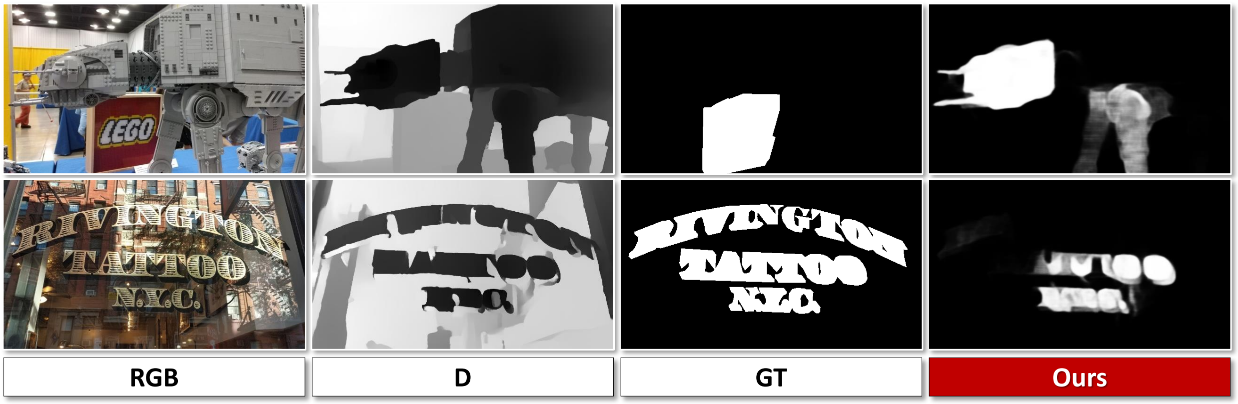


Fig 6 Demonstration of some representative failure cases.

fusion to address the issue of weakened local information during the capture of long-range dependencies. A spatial info sensing module is integrated to refine the multimodal features to enhance compatibility further.

The experimental results on four public datasets validate the superiority of the proposed NCFE-Net, *e.g.*, an average increase of 0.4%, 0.3%, 0.7%, and 0.2% in terms of the S-measure metric of the four public datasets. It competes with and surpasses state-of-the-art methods in terms of accuracy and efficiency. This demonstrates the potential of non-convolutional approaches in salient object detection, with NCFE-Net striking a favorable balance between performance and speed compared to CNN-based and Transformer-based methods. Overall, this approach opens up promising avenues for future research in salient object detection and holds potential for application in other computer vision tasks.

Code, Data, and Materials Availability at: https://github.com/xl0312/InoSal

References

1 W. Wang, J. Shen, R. Yang, *et al.*, "Saliency-aware video object segmentation," *IEEE TPMAI* (2018).

2 X. Shen, J. Yang, C. Wei, *et al.*, “Dct-mask: Discrete cosine transform mask representation for instance segmentation,” in *IEEE CVPR*, (2021).

3 X. Yuan, A. Kortylewski, Y. Sun, *et al.*, “Robust instance segmentation through reasoning about multi-object occlusion,” in *IEEE CVPR*, (2021).

4 P. Zhang, W. Liu, D. Wang, *et al.*, “Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps,” *PR* (2020).

5 J. Cai, M. Xu, W. Li, *et al.*, “Memot: multi-object tracking with memory,” in *IEEE CVPR*, (2022).

6 Z. Cao, Z. Huang, L. Pan, *et al.*, “Tctrack: Temporal contexts for aerial tracking,” in *IEEE CVPR*, (2022).

7 C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” *IEEE TIP* (2010).

8 X. Zhang and X. Wu, “Attention-guided image compression by deep reconstruction of compressive sensed saliency skeleton,” in *IEEE CVPR*, (2021).

9 J. Shi, N. Xu, Y. Xu, *et al.*, “Learning by planning: Language-guided global image editing,” in *IEEE CVPR*, (2021).

10 G. Hu and C. Saeli, “Scale-invariant salient edge detection,” in *IEEE ICIP*, (2021).

11 Y. Piao, Z. Rong, M. Zhang, *et al.*, “A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection,” in *IEEE CVPR*, (2020).

12 J. Zhang, D.-P. Fan, Y. Dai, *et al.*, “Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders,” in *IEEE CVPR*, (2020).

13 K. He, X. Zhang, S. Ren, *et al.*, “Deep residual learning for image recognition,” in *IEEE CVPR*, (2016).

14 K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556* (2014).

15 N. Carion, F. Massa, G. Synnaeve, *et al.*, “End-to-end object detection with transformers,” in *ECCV*, (2020).

16 D. Li, J. Hu, C. Wang, *et al.*, “Involution: Inverting the inherence of convolution for visual recognition,” in *IEEE CVPR*, (2021).

17 C. Zhu and G. Li, “A multilayer backpropagation saliency detection algorithm and its applications,” *Mul. Too. and App.* (2018).

18 C. Zhu, G. Li, W. Wang, *et al.*, “An innovative salient object detection using center-dark channel prior,” in *IEEE ICCVW*, (2017).

19 L. Wu, Z. Liu, H. Song, *et al.*, “Rgbd co-saliency detection via multiple kernel boosting and fusion,” *Mul. Too. AND. App.* (2018).

20 H. Song, Z. Liu, Y. Xie, *et al.*, “Rgbd co-saliency detection via bagging-based clustering,” *IEEE SPL* (2016).

21 J. Ren, X. Gong, L. Yu, *et al.*, “Exploiting global priors for rgb-d saliency detection,” in *IEEE CVPRW*, (2015).

22 X. Dan, H. Shuheng, and Z. Xin, “Spatial-aware global contrast representation for saliency detection,” *Turkish Journal of Electrical Engineering and Computer Sciences* (2019).

23 J. Shang, Y. Liu, H. Zhou, *et al.*, “Moving object properties-based video saliency detection,” *JEI* (2021).

24 Y. Gao, S. Dai, W. Ji, *et al.*, “Low saliency crack detection based on improved multimodal object detection network: an example of wind turbine blade inner surface,” *JEI* (2023).

25 M. Du, X. Wu, W. Chen, *et al.*, “Exploiting multiple contexts for saliency detection,” *JEI* (2016).

26 W. Li, S. Feng, H.-P. Guan, *et al.*, “Video saliency detection based on low-level saliency fusion and saliency-aware geodesic,” *JEI* (2019).

27 X. Wang, S. Shen, and C. Ning, “Visual saliency detection based on in-depth analysis of sparse representation,” *JEI* (2018).

28 C. Liu, D. Zhang, and X. Zhao, “Multitask saliency detection model for synthetic aperture radar (sar) image and its application in sar and optical image fusion,” *JEI* (2018).

29 Y. Zhou, Q. Li, Y. Ma, *et al.*, “Salient object detection via joint perception of region-level spatial distribution and color contrast,” *JEI* (2021).

30 P. Guo, Y. Wang, L. Wang, *et al.*, “Image saliency detection based on regional label fusion,” in *ICIEIS*, Y. Zhou and Z. Chen, Eds. (2022).

31 Y. Li and X. Mou, “Saliency detection based on structural dissimilarity induced by image quality assessment model,” *JEI* .

32 Y. Shi, Y. Yi, K. Zhang, *et al.*, “Multiview saliency detection based on improved multimanifold ranking,” *JEI* (2014).

33 X. Zhou, H. Wen, R. Shi, *et al.*, “Fanet: Feature aggregation network for rgbd saliency detection,” *SPIC* (2022).

34 Z. Zhang, Z. Lin, J. Xu, *et al.*, “Bilateral attention network for rgb-d salient object detection,” *IEEE TIP* (2021).

35 Y. Pang, L. Zhang, X. Zhao, *et al.*, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *ECCV*, (2020).

36 P. Sun, W. Zhang, H. Wang, *et al.*, "Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *IEEE CVPR*, (2021).

37 D. Misra, T. Nalamada, A. U. Arasanipalai, *et al.*, "Rotate to attend: Convolutional triplet attention module," in *IEEE WACV*, (2021).

38 W. Zhang, Y. Jiang, K. Fu, *et al.*, "Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection," in *IEEE ICME*, (2021).

39 K. Fu, D.-P. Fan, G.-P. Ji, *et al.*, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *IEEE CVPR*, (2020).

40 X. Zhou, H. Fang, Z. Liu, *et al.*, "Dense attention-guided cascaded network for salient object detection of strip steel surface defects," *IEEE TIM* (2021).

41 H. Song, Z. Liu, H. Du, *et al.*, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE TIP* (2017).

42 J. Guo, T. Ren, and J. Bei, "Salient object detection for rgb-d image via saliency evolution," in *IEEE ICME*, (2016).

43 R. Cong, J. Lei, C. Zhang, *et al.*, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE SPL* (2016).

44 Z. Liu, S. Shi, Q. Duan, *et al.*, "Salient object detection for rgb-d image by single stream recurrent convolution neural network," *Neurocomputing* (2019).

45 X. Wang, S. Li, C. Chen, *et al.*, "Data-level recombination and lightweight fusion scheme for rgb-d salient object detection," *IEEE TIP* (2020).

46 C. Chen, J. Wei, C. Peng, *et al.*, “Depth-quality-aware salient object detection,” *IEEE TIP* (2021).

47 H. Chen and Y. Li, “Progressively complementarity-aware fusion network for rgb-d salient object detection,” in *IEEE CVPR*, (2018).

48 G. Li, Z. Liu, and H. Ling, “Icnet: Information conversion network for rgb-d based salient object detection,” *IEEE TIP* (2020).

49 D.-P. Fan, Z. Lin, Z. Zhang, *et al.*, “Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks,” *IEEE TNNLS* (2021).

50 G. Li, Z. Liu, and H. Ling, “Icnet: Information conversion network for rgb-d based salient object detection,” *IEEE TIP* (2020).

51 C. Chen, J. Wei, C. Peng, *et al.*, “Depth-quality-aware salient object detection,” *IEEE TIP* (2021).

52 M. Zhang, W. Ren, Y. Piao, *et al.*, “Select, supplement and focus for rgb-d saliency detection,” in *IEEE CVPR*, (2020).

53 A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *NIPS*, (2017).

54 Z. Liu, Y. Lin, Y. Cao, *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *IEEE ICCV*, (2021).

55 Z. Liu, H. Hu, Y. Lin, *et al.*, “Swin transformer v2: Scaling up capacity and resolution,” in *IEEE CVPR*, (2022).

56 Z. Liu, Y. Wang, Z. Tu, *et al.*, “Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network,” in *ACM MM*, (2021).

57 L. Tang and B. Li, "Cosformer: Detecting co-salient object with transformers," *arXiv preprint arXiv:2104.14729* (2021).

58 S. Ren, Q. Wen, N. Zhao, *et al.*, "Unifying global-local representations in salient object detection with transformer," *arXiv preprint arXiv:2108.02759* (2021).

59 Y. Wang, X. Jia, L. Zhang, *et al.*, "Transformer-based network for rgb-d saliency detection," *arXiv preprint arXiv:2112.00582v1* (2021).

60 R. Ju, L. Ge, W. Geng, *et al.*, "Depth saliency based on anisotropic center-surround difference," in *IEEE ICIP*, (2014).

61 H. Peng, L. Bing, W. Xiong, *et al.*, "Rgbd salient object detection: A benchmark and algorithms," in *ECCV*, (2014).

62 Y. Niu, Y. Geng, X. Li, *et al.*, "Leveraging stereopsis for saliency analysis," in *IEEE CVPR*, (2012).

63 D.-P. Fan, M.-M. Cheng, Y. Liu, *et al.*, "Structure-measure: A new way to evaluate foreground maps," in *IEEE ICCV*, (2017).

64 R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *IEEE CVPR*, (2014).

65 J. Zhao, C. Y, D.-P. Fan, *et al.*, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE CVPR*, (2019).

66 Y. Piao, W. Ji, J. Li, *et al.*, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE ICCV*, (2019).

67 M. Zhang, S. Fei, J. Liu, *et al.*, "Asymmetric two-stream architecture for accurate rgb-d saliency detection," in *ECCV*, (2020).

68 Y. Zhai, D.-P. Fan, J. Yang, *et al.*, "Bifurcated backbone strategy for rgb-d salient object detection," *IEEE TIP* (2021).

69 C. Li, R. Cong, S. Kwong, *et al.*, "Asif-net: Attention steered interweave fusion network for rgb-d salient object detection," *IEEE TCYB* (2021).

70 M. Song, W. Song, G. Yang, *et al.*, "Improving rgb-d salient object detection via modality-aware decoder," *IEEE TIP* (2022).

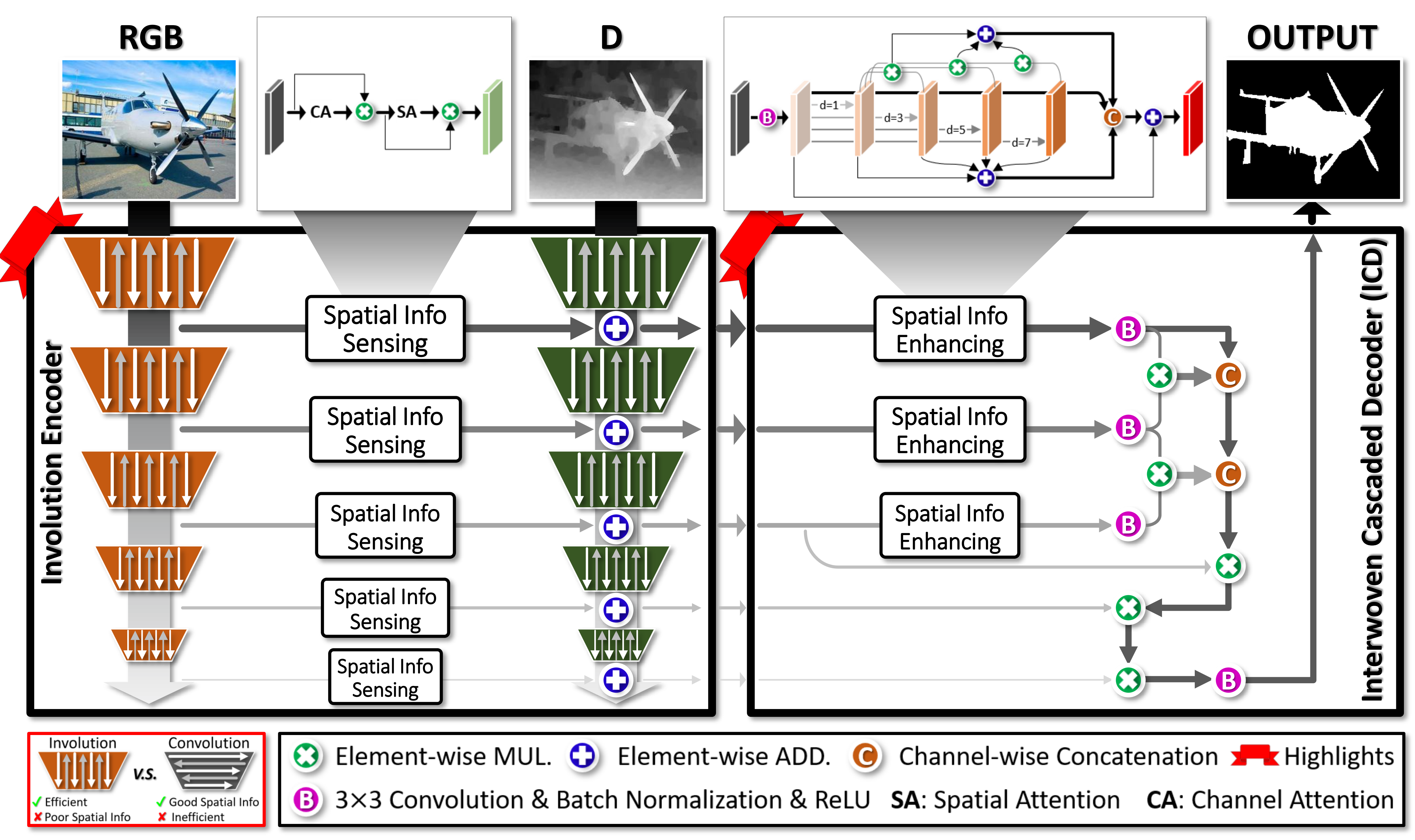
71 Y.-H. Wu, Y. Liu, J. Xu, *et al.*, "Mobilesal: Extremely efficient rgb-d salient object detection," *IEEE TPAMI* (2022).

72 X. Fang, J. Zhu, X. Shao, *et al.*, "Grouptransnet: Group transformer network for rgb-d salient object detection," *arXiv preprint arXiv:2203.10785v1* (2022).

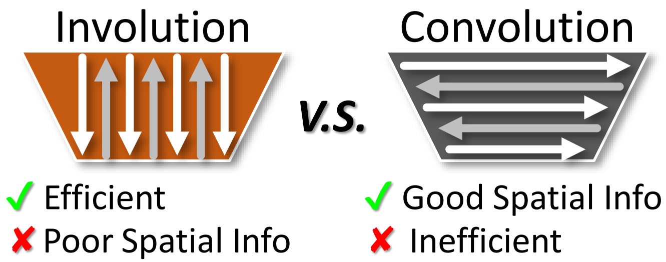
73 Y. Pang, X. Zhao, L. Zhang, *et al.*, "Caver: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE TIP* (2023).

List of Figures

List of Tables

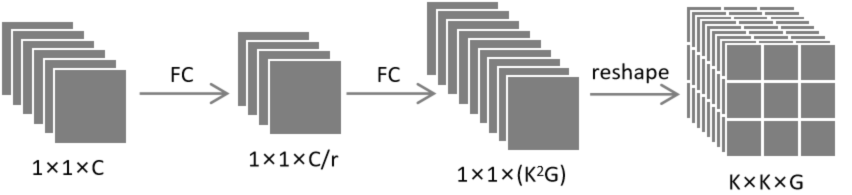


RGB
D
OUTPUT
CA
SA
d=1
d=3
d=5
d=7
Involution Encoder
Spatial Info Sensing
Spatial Info Sensing
Spatial Info Sensing
Spatial Info Sensing
Spatial Info Sensing
Interwoven Cascaded Decoder (ICD)
Spatial Info Enhancing
Spatial Info Enhancing
Spatial Info Enhancing
Involution
Convolution
v.s.
✓ Efficient
✗ Poor Spatial Info
✓ Good Spatial Info
✗ Inefficient
Element-wise MUL.
Element-wise ADD.
Channel-wise Concatenation
Highlights
3×3 Convolution & Batch Normalization & ReLU
SA: Spatial Attention
CA: Channel Attention

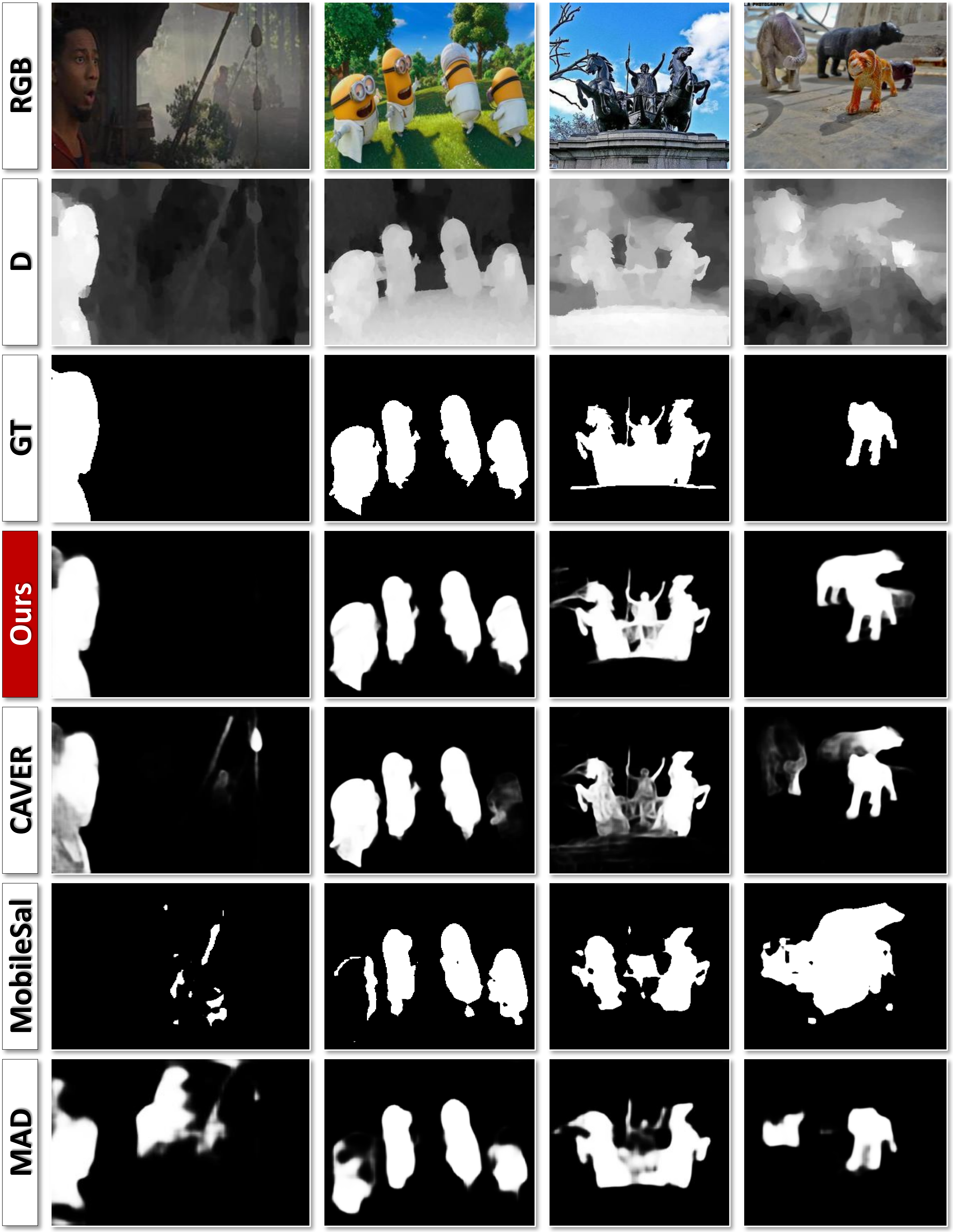


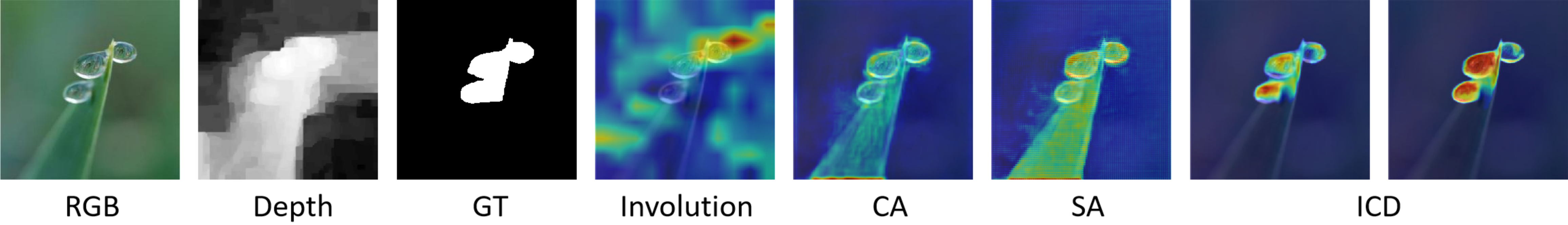
Element-wise MUL. Element-wise ADD. Channel-wise Concatenation Highlights

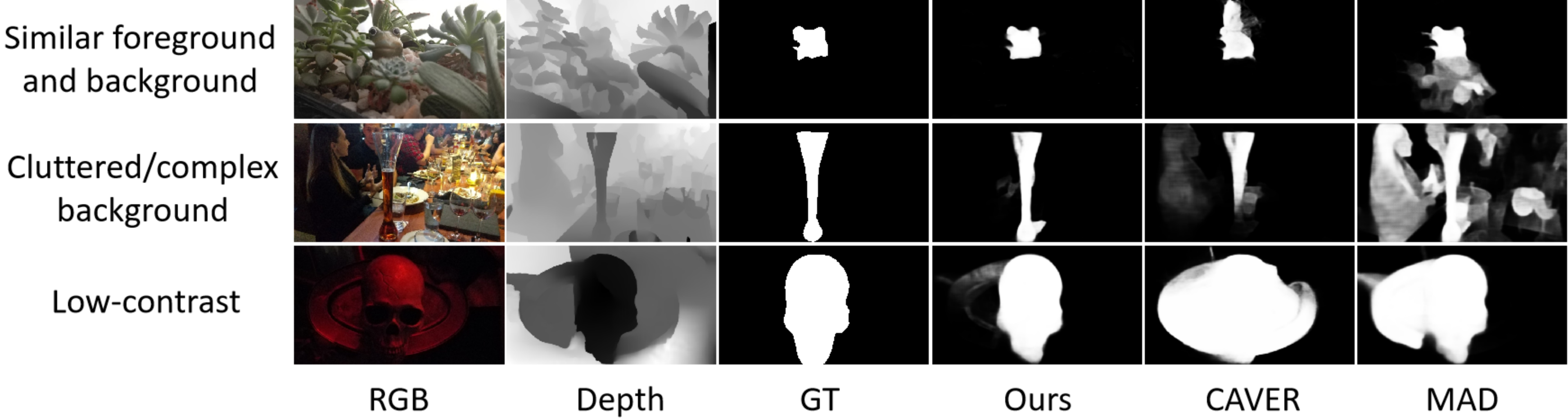
3×3 Convolution & Batch Normalization & ReLU **SA**: Spatial Attention **CA**: Channel Attention

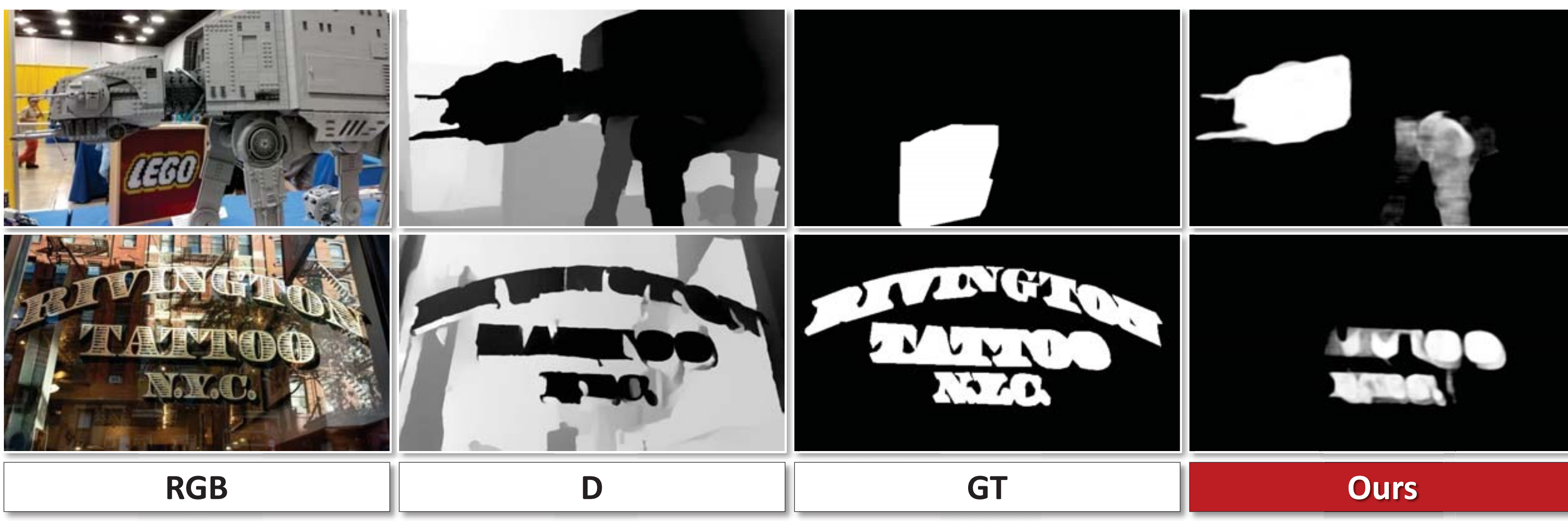


1×1×C
FC
1×1×C/r
FC
1×1×(K²G)
reshape
K×K×G


RGB
D
GT
Ours
CAVER
MobileSal
MAD


RGB
Depth
GT
Involution
CA
SA
ICD


Similar foreground and background
Cluttered/complex background
Low-contrast
RGB
Depth
GT
Ours
CAVER
MAD


LEGO
RIVINGTON
TATTOO
N.Y.C.
RGB
D
GT
Ours