2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

DMKAN-Net: A Dual-Modal Fusion and KAN Decoder Network for RGB-D Salient Object Detection

Qingbo Xue[0009-0008-1836-428X], Panpan Zheng [0009-0003-2934-6339]
and Liejun Wang(✉)[0000-0003-0210-2273]

School of Computer Science and Technology, Xinjiang University, 830017 Urumqi, China
`wljxju@xju.edu.cn`

Abstract. At present, most RGB-D saliency target detection algorithms (SOD) are committed to investing a lot of computing power in the encoding and feature fusion stages. Admittedly, this strategy brings the performance improvement, but it also ignores the feature recovery ability and the fitting ability in the decoding stage. Therefore, we designed a dual-modal fusion and KAN decoder network (called DMKAN-Net) to better implement the RGB-D SOD task. This network structure has only three parts, dual-stream encoder, dual-modal converter and KAN decoder. Among them, the dual-stream encoder adopts the Swin Transformer encoder, which is mainly used to extract the multilevel and global features in the RGB and depth images. In the dual-modal fusion section, we design a dual-modal feature fusion module to capture the channel information and spatial information in different modes and fuse it. KAN decoder is the decoder mainly composed of KAN module, which uses nonlinear and learnable activation function to better recover and predict saliency targets. Moreover, experiments performed on five benchmark datasets show that our method achieves competitive results.

Keywords: RGB-D, SOD, Dual-modal, KAN.

1 Introduction

The purpose of salient object detection [1] (SOD) is to identify and segment the most attractive target or area in the image. It has also been widely used in multiple fields of computer vision, such as object detection [2], visual tracking [3], image retrieval [4], and instance segmentation [5]. However, RGB images show complex scenes, color distortion, and messy background phenomena, which poses a challenge for saliency object detection in complex scenes. Therefore, thanks to the ability of low-cost depth sensors that can easily obtain depth information in the scene, the RGB-D salient object detection [6] (RGB-D SOD) algorithm has attracted wide attention from researchers.

The traditional RGB-D SOD algorithm usually detects saliency targets using manual features and different prior information, but it cannot be well generalized to all images. Due to the rapid development of deep learning algorithms, its efficient feature extraction ability makes it the mainstream algorithm in the field of computer vision. Qu et al. [7] proposed the first deep learning method for RGB-D image saliency object detection in 2017, and achieved significant improvement. Since then, RGB-D SOD algorithm based on deep learning has become the mainstream of the RGB-D SOD algorithm.

Dual-Modal information fusion has always been a central issue in RGB-D SOD. Compared with other fusion methods, feature-level fusion can often receive good results. Such as Wu et al. [8] proposed a multiscale multilevel encoder fusion scheme with cross-domain supervision and decoder fusion that utilizes channel dependence. Cong et al. [9] introduce the progressive attention-guided integration unit and importance-gating fusion, which integrates RGB and depth functions in the encoder and decoder stages, respectively. And Zhang et al. [10] propose a bidirectional transmission and selection module that enables RGB and depth to correct and optimize each other at the encoder stage.

How to better carry out dual-modal feature fusion is on the one hand, and how to better apply the fusion features to the prediction stage, that is, the decoding stage, is on the other hand, many RGB-D saliency object detection only use some simple convolution combination as the decoding stage, this approach ignores the importance of decoding.

In order to solve the above problems, we innovatively proposed a dual-modal fusion and KAN decoding network, making full use of the different information of dual-modal for feature fusion to obtain more semantic information, and then introduce the KAN module into the SOD task, using the KAN module can be non-linear and learnable activation function to have stronger approximation ability, to improve the prediction ability and interpretability of the network.

Our main contributions are as follows.

- We propose a dual-modal fusion and KAN decoder network which both effectively incorporates the features of the dual-modal and considers the interpretability and validity of the prediction phase, proposing the KAN decoder. Experiments on five publicly available datasets show that our model has this excellent performance.
- In the feature fusion stage, we propose a two-modal feature fusion module to capture more spatial and channel information by passing the RGB and Depth features through the attention module, respectively, for the final saliency prediction phase.
- In the decoding stage, we innovatively introduced the KAN decoder, which uses the good approximation ability and interpretability of the KAN module to obtain a good significance prediction map.

2 Related works

2.1 Dual-modal feature fusion method in salient object detection

RGB-D feature fusion methods are widely used in salient object detection. RGB images provide rich color and texture information, while depth images complement the spatial structure of objects. By effectively fusing the features of both, the accuracy and robustness of the task can be improved. Common RGB-D feature fusion methods include the following.

Late Fusion

Late fusion methods first independently extract features from the RGB and depth images, and then fuse the extracted features, typically at higher layers of the network. Piao et al. [11] proposed using independent RGB and depth branches to extract features, and employing a multi-scale recurrent attention mechanism to adaptively weight and fuse the features at the decision stage. Han et al. [12] proposed a multi-view fusion strategy, where RGB and depth CNNs are trained separately, and the prediction results are fused later to enhance the cross-modal information complementarity.

Cascade Fusion

Cascade fusion in RGB-D salient object detection typically refers to the gradual and staged fusion of RGB and depth image information, progressively improving the saliency prediction results. AFNet [13], CPNet [14], and HiDANet [8] are some typical networks that optimize the fusion of RGB-D information through different cascade fusion strategies. For example, AFNet introduces an attention fusion mechanism to gradually fuse RGB and depth features at different stages, using both global and local attention to selectively weight and fuse important feature information. CPNet performs pixel-level cross-modal fusion, progressively fusing RGB and depth maps at multiple layers to fine-tune the saliency score of each pixel. The network gradually optimizes the prediction results at each layer by learning the pixel-level relationships between the two modalities. HiDANet, on the other hand, adopts a high-dimensional attention mechanism, which enhances the representation of both RGB and depth information by progressively fusing them in the feature space. It dynamically adjusts the contributions of the two modalities at multiple stages through the high-dimensional attention mechanism. These methods demonstrate that cascade fusion not only integrates features step by step but also performs fine-grained optimization at each stage to fully leverage the complementary strengths of RGB and depth maps.

2.2 Kolmogorov–Arnold Networks (KANs)

Kolmogorov Arnold theorem shows that any continuous function can be expressed as a combination of continuous unary functions of finite variables. Hornik et al. [15] further proved that the feedforward neural network has the general approximation ability. Based on Kolmogorov Arnold theorem, scholars proposed a new neural network architecture, Kolmogorov Arnold networks (KANs) [16]. KANs consists of a series of Kolmogorov Arnold layers in series, and each layer contains a set of one-dimensional activation functions that can be learned. This architecture performs well in approaching high-dimensional complex functions, and has shown strong performance in many applications.

KANs is especially famous for its strong theoretical interpretability and interpretability. Huang et al. [17] analyzed the optimization characteristics and convergence behavior of KANs, and confirmed its excellent approximation ability and generalization performance. In addition, Liang et al. [18] introduced the depth KAN model and successfully applied it to image classification and other tasks, while Xing et al. [19] explored the application of KANs in time series prediction and control.

3 Proposed Method

3.1 Overall Architecture

In this work, as illustrated in **Fig. 1** we propose a holistic network architecture based on a Transformer that integrates dual-modal feature fusion and decoding stages. The framework comprises three principal components: (1) a siamese Swin Transformer serving as the encoder; (2) four dual-modal feature fusion attention modules (DMFAMs); (3) a three-stage decoding cascade primarily constructed using KAN Blocks. Initially, the siamese backbone network extracts multi-level features $F_i^{r/d},\ i \in \{1,2,3,4\}$ from paired RGB and depth images through parallel hierarchical processing. These multi-scale features are subsequently fed into four cascaded DMFAMs for progressive cross-modal interaction and feature alignment. Finally, the fused features $F_i,\ i \in \{1,2,3,4\}$ at each hierarchy are decoded through a KAN Block-based decoder that systematically aggregates low-level spatial details to enhance semantic coherence, ultimately generating refined saliency maps. Detailed architectural specifications and implementation particulars of each component will be elaborated in subsequent sections.

3.2 Siamese Swin Transformer Encoder

High-quality feature extraction constitutes a fundamental challenge in salient object detection (SOD). While conventional SOD frameworks predominantly employ Convolutional Neural Networks (CNNs) for feature representation, recent advances have demonstrated the superiority of Transformer architectures due to their enhanced feature extraction capacity and global contextual modeling. The Swin Transformer [20], in particular, addresses computational efficiency constraints through its innovative shifted window mechanism, which achieves global interaction modeling while maintaining linear computational complexity relative to input resolution. Motivated by these advantages, we adopt a siamese Swin Transformer-B backbone (selected for optimal performance-efficiency trade-offs) to extract hierarchical multi-scale features from paired RGB and depth modalities.

As depicted in **Fig. 1**, our pipeline begins by converting depth maps from single-channel to three-channel format to ensure dimensional consistency with RGB inputs. Subsequently, the dual-modal images are partitioned into non-overlapping patches via a patch partitioning operation. These patches are then fed into a series of cascaded Swin Transformer blocks to extract hierarchical features across four stages, generating multi-level representations $F_i^{r/d},\ i \in \{1,2,3,4\}$ for both RGB and depth modalities. The encoder comprises four hierarchical stages, each containing a patch merging layer (replaced by a patch embedding layer in the first stage) and multiple stacked Swin Transformer blocks. This architecture yields complementary multi-scale representations from both RGB and depth streams, with stage-wise feature maps progressively capturing semantic granularity from local textures to global context.

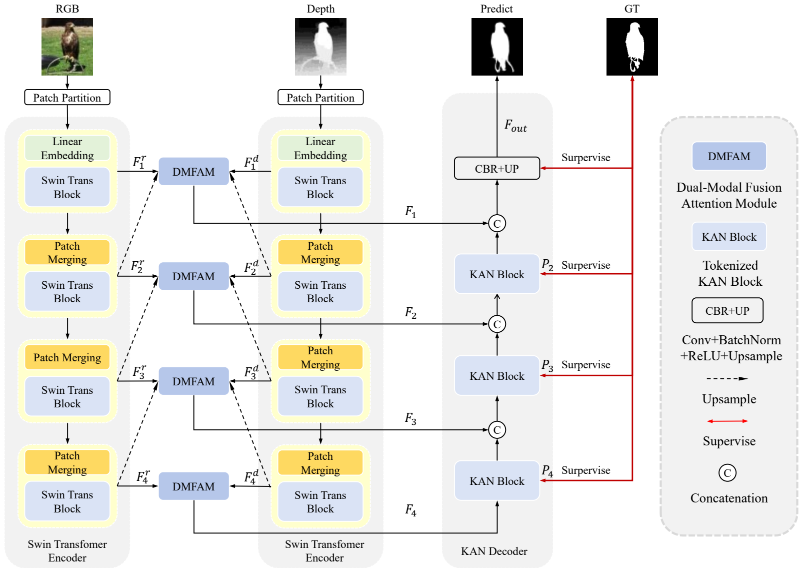

Fig. 1. Our proposed DMKAN-Net.

3.3 Dual-Modal Fusion Attention Module

Given that RGB features encompass substantial texture information, whereas depth features harbor richer spatial positional data, the challenge lies in how to efficiently harness the information embedded within both RGB and depth features to facilitate feature fusion, which constitutes a pivotal concern in RGB-D SOD tasks. For this purpose, we have designed a Dual-Modal Fusion Attention Module (DMFAM) to effectively integrate RGB features and depth features. Furthermore, we have observed that some SOD methods exhibit poor performance when detecting smaller objects. We believe that this is attributed to the insufficient contextual information contained within single-level features, which makes it difficult to handle objects of various sizes. Therefore, we integrate current-level features with enlarged high-level features, leveraging the guidance of high-level semantic information to obtain rich multi-scale contextual information and enhance the detection capability for targets of different scales.

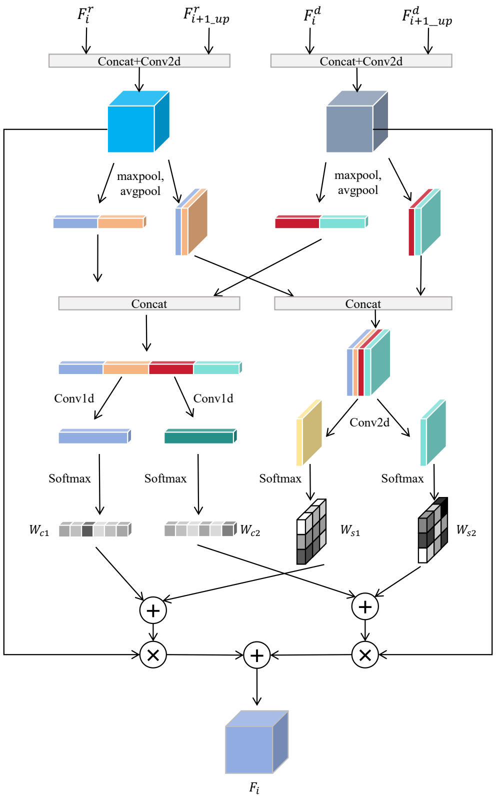

Fig. 2. The proposed Dual-Modal Fusion Attention Module.

As illustrated in **Fig. 2**, it concurrently utilizes both spatial attention and channel attention mechanisms to identify significant portions within the input features. Furthermore, it performs cross-level fusion and dual-modal fusion to determine the important components between features of different levels and modalities.

We first upsample the high-level features $F_{i+1}^{r/d}$, $i \in \{1,2,3\}$ to match the size of $F_i^{r/d}$, $i \in \{1,2,3\}$ and concatenate them with $F_i^{r/d}$ (when $i=4$, we concatenate $F_4^{r/d}$ with itself). Subsequently, the concatenated features are passed through a convolutional layer to obtain multi-scale features $\tilde{F}_i^{r/d}$ .The specific process and formulation are as follows:

$$\tilde{F}_i^{r/d} = Conv_{3x3}(Cat(Up(F_{i+1}^{r/d}), F_i^{r/d})). \tag{1}$$

where $Up(\cdot)$ denotes upsampling, $Cat(\cdot)$ represents concatenation, and $Conv_{3x3}(\cdot)$ stands for 3×3 convolution operation. Subsequently, the obtained multi-scale features are separately fed into channel branch and spatial branch. In the channel branch, the input multi-scale features undergo global pooling across the spatial dimensions to aggregate spatial information. The specific process can be formulated as follows:

$$S_c = Cat(Avg(\tilde{F}_i^r), Max(\tilde{F}_i^r), Avg(\tilde{F}_i^d), Max(\tilde{F}_i^d)). \tag{2}$$

where S_c denotes the aggregated spatial features, $Avg(\cdot)$ and $Max(\cdot)$ represent the operations of global average pooling and global max pooling along the spatial dimensions, respectively.

The aggregated spatial features are separately passed through two 1D convolutional layers to determine the weights for the input dual-modal channels. Subsequently, the obtained dual-modal channel weights are passed through a $Softmax(\cdot)$ function to ensure their weights sum to 1. This means that by comparing the dual-modal weights, we can determine which has a higher value, thereby identifying the important parts between dual-modal features in the channel dimension. The specific process can be represented as follows:

$$W_{c1}, W_{c2} = \sigma(Conv_1(S_c)), \sigma(Conv_2(S_c)). \tag{3}$$

where W_{c1}, W_{c2} denotes the dual-modal weights, σ represents the $Softmax(\cdot)$ function and $Conv_1$, $Conv_2$ stands for the 1D-convolution operation. While the dual-modal spatial weights W_{s1}, W_{s2} are also determined in a similar manner, with the exception that the $Avg(\cdot)$ and $Max(\cdot)$ are performed along the channel dimensions instead. By summarizing the four weights obtained, we can derive the dual-modal weights, which identify the important parts within the dual-modal features. Finally, multiplying and summing the dual-modal weights with the dual-modal features effectively fuse the dual-modal features. The fused features can be represented as

$$F_i = (W_{c1} + W_{s1}) \odot \tilde{F}_i^r + (W_{c2} + W_{s2}) \odot \tilde{F}_i^d. \tag{4}$$

where F_i represents the fused features, and $\odot$ denotes element-wise multiplication. When the sum of the dual-modal weights equals 1, the effective parts between the dual-modal features are preserved, while the useless parts are discarded, thereby achieving effective feature fusion.

3.4 KAN Decoder

The decoding stage of salient object detection primarily involves using convolutional layers and upsampling to restore the image resolution, followed by the application of activation functions to generate the saliency map. Inspired by U-KAN [21], we have designed the KAN decoder to perform the decoding task in salient object detection.

The overall structure of the KAN Decoder is shown in **Fig. 1**. During the decoding stage, we use three KAN modules to decode the features F_4, F_3 and F_2, respectively. The structure of a single KAN module is illustrated in **Fig. 3** where **Fig. 3**(a) shows the serialized KAN module, and figure (b) presents the original implementation of the KAN layer.

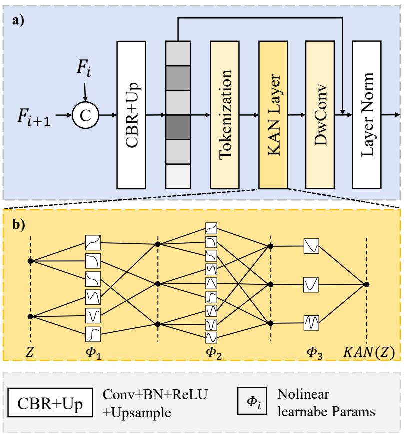

Fig. 3. The proposed Tokenized KAN Decoder Block, a) means KAN Block b) means KAN Layer.

Similar to MLP, the K-layer KAN Layer can be viewed as a nested structure of K individual KAN functions. For a given input Z the output Z' of the KAN layer can be described by the following equation:

$$Z' = KAN(Z) = (\Phi_{K-1} \circ \Phi_{K-2} \circ \cdots \circ \Phi_1 \circ \Phi_0)Z. \tag{5}$$

where Φ_i represents the i -th function of the entire KAN layer. Each KAN function has an input of dimension n_{in} and output of dimension n_{out} . Then, it is represented by n learnable functions, which can be expressed as:

$$\begin{aligned} &\Phi = \{\phi_{p,q}\}. \\ &p = 1, 2, \cdots, n_{in}. \\ &q = 1, 2, \cdots, n_{out}. \end{aligned} \tag{6}$$

The K -th layer in the KAN Layer can be expressed by the following formula:

$$\Phi_K = \begin{pmatrix} \phi_{K,1,1}(\cdot) & \phi_{K,1,2}(\cdot) & \cdots & \phi_{K,1,n_{out}}(\cdot) \\ \phi_{K,2,1}(\cdot) & \phi_{K,2,2}(\cdot) & \cdots & \phi_{K,2,n_{out}}(\cdot) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{K,n_{in},1}(\cdot) & \phi_{K,n_{in},2}(\cdot) & \cdots & \phi_{K,n_{in},n_{out}}(\cdot) \end{pmatrix}. \tag{7}$$

Overall, the KAN Layer approximates complex nonlinear functions through a series of learnable univariate simple functions.

In the Tokenized KAN Block, after receiving the input features, the module first performs upsampling, then reshapes the features into a flattened 2D patch sequence through tokenization for further processing by the KAN Layer. In the final step of the KAN Layer, the features are restored to their original 2D form and refined through depthwise separable convolutions to adjust the feature details. Finally, the initial input features are added to the current features via a shortcut branch, and the features of the current layer are adjusted through Layer Normalization.

3.5 Loss Function

In this paper, we adopt a deep supervision strategy to generate saliency prediction maps F_{out} and $P_i, i \in \{2,3,4\}$ from KAN Block features. We employ a hybrid loss consisting of the Binary Cross-Entropy (BCE) loss [22] and the Intersection over Union (IoU) loss [23] to supervise these maps. The hybrid loss can accurately supervise at the pixel level and distinguish between foreground and background regions. Additionally, with the help of the deep supervision strategy, the recognition capability of the network can be enhanced, and the convergence speed can be accelerated.

The BCE loss is defined as:

$$L_{BCE} = -\sum_{x=1}^{H}\sum_{y=1}^{W}[G(x,y)log(P(x,y)) + (1-G(x,y))log(1-P(x,y))]. \tag{8}$$

where W and H represent the width and height of the image, respectively.

The IoU loss is defined as:

$$L_{IoU} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} P(x,y)G(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W}[P(x,y)+G(x,y)-P(x,y)G(x,y)]}. \tag{9}$$

The total loss L of the model is defined as:

$$L = \sum_{i=2}^{4}[L_{BCE}^{i}(P_i,G) + L_{IoU}^{i}(P_i,G)] + L_{BCE}^{1}(F_{out},G) + L_{IoU}^{1}(F_{out},G). \tag{10}$$

where P_i represent the feature generated by KAN Block, F_{out} represent the output of the DMKAN-Net.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets

In order to comprehensively evaluate the performance of our proposed network, we conducted extensive experiments on five challenging RGB-D SOD datasets. The NLPR [24] contains 1,000 images with single or multiple salient objects. The NJU2K [25] provides 2,003 stereo image pairs of varying resolutions. The SIP [26] includes 1,000 high-resolution images of public figures. The LFSD [27] consists of 100 images captured by a Lytro camera, featuring multiple small objects and complex backgrounds. STERE [28] contains 1,000 stereo images with masks highlighting the objects of interest.

In order to ensure a fair comparison, we adopt the same training dataset as used by Ji et al. [29] and Zhao et al. [30], which comprises 1,485 images from the NJU2K dataset, 700 images from the NLPR dataset, and 800 images from the DUT Piao et al. [11] dataset, totaling 2,985 samples for training our model.

Evaluation Metrics

We employ four widely used evaluation metrics to assess our model, namely E-measure (E_ξ) [31], S-measure (S_m) [32], F-measure (F_β) [33], and Mean Absolute Error (MAE) [34]. Specifically, E-measure (E_ξ) is used to measure both local pixel-level and global image-level matching information. S-measure (S_m) evaluates the region-aware and object-aware structural similarity of the spatial layout between the saliency map and the ground truth. F-measure (F_β) is the weighted harmonic mean of precision and recall, which can be used to evaluate the overall performance. MAE measures the average absolute difference per pixel between the saliency map and the ground truth. In our experiments, maximum values are adopted for E-measure and F-measure.

Experiment Details

During both training and test phases, all input RGB and depth images are spatially resampled to a uniform resolution of 384×384 pixels. The depth images are duplicated into three channels to match the format of RGB images. During the training process, we employ data augmentation techniques such as random flipping, rotation, and border cropping. We initialize the parameters of the backbone network using a pre-trained Swin-B model, while the remaining parameters are initialized to PyTorch's default settings. We use the Adam optimizer [35] to train our network with a batch size of 8, an initial learning rate of 5e-5, and the learning rate is divided by 10 every 100 epochs. Our model is trained on a machine with a single NVIDIA A40 GPU. The model converges within 200 epochs.

4.2 Comparisons with the State-of-the-art

To thoroughly validate the effectiveness of our proposed model, we compared it with 13 existing deep learning-based RGB-D salient object detection methods, including AFNet [13], CFIDNet [36], ICNet [37], DMRA [11], UCNet [38], JLDCF [39], DANet [40], BBS-Net [41], BTS-Net [42], CATNet [43], AMINet [44], DCMNet [45], and HRTransNet [46]. To ensure a fair comparison, we used the saliency maps provided by the authors. If these were not available, we generated the saliency maps using the source code and model files provided by the authors.

Quantitative experiment

Table 1 presents the quantitative evaluation results for four evaluation metrics in five datasets, clearly showing the exceptional performance of our proposed DMKAN-Net method. For the metrics maxE and maxF, our method outperformed others on all datasets except for LFSD. Overall, our method achieved the best results in most of the metrics, which can verifiy the effectiveness and generalization of the proposed algorithm.

Table 1. Comparison of evaluation results on four evaluation metrics - MAE, max F-measure (maxF), max E-measure (maxE), and S-measure (S) - across five datasets. The arrow ↑ indicates that a higher value is better, while ↓ signifies that a lower value is preferable. The best two results are shown in red, and blue fonts, respectively. '-' indicates the code or result is not available.

Datasets	Metrics	DA Net	JL-DCF	UC Net	BBS Net	DMRA	IC Net	CFID Net	AF Net	BTS Net	CAT Net	AMI Net	DCM Net	HRTransNet	Ours
Parameters(M)		**106.7**	-	-	**49.77**	**60.31**	-	-	**258.1**	-	**262.6**	**199.1**	-	**58.9**	**229.4**
Flops(G)		-	-	-	**31.40**	**126.29**	-	-	**130.0**	-	**341.8**	**124.7**	-	**17.1**	**124.4**
LFSD	MAE↓	0.083	0.081	0.067	0.072	0.074	0.071	0.071	0.056	0.071	0.051	0.056	0.064	-	0.055
	maxF↑	0.846	0.854	0.863	0.858	0.858	0.870	0.865	0.888	0.873	0.894	0.883	0.867	-	0.885
	maxE↑	0.886	0.887	0.905	0.900	0.905	0.903	0.903	0.923	0.906	0.908	0.906	0.906	-	0.920
	S↑	0.845	0.849	0.864	0.864	0.845	0.868	0.869	0.890	0.867	0.894	0.871	-	-	0.915
NJU2K	MAE↓	0.047	0.039	0.035	0.035	0.049	0.052	0.038	0.032	0.037	0.025	0.035	0.036	0.026	0.025
	maxF↑	0.893	0.915	0.910	0.920	0.892	0.891	0.915	0.928	0.902	0.929	0.912	0.899	0.928	0.940
	maxE↑	0.936	0.951	0.949	0.949	0.937	0.926	0.946	0.958	0.942	0.933	0.928	0.920	0.931	0.963
	S↑	0.897	0.913	0.911	0.921	0.889	0.894	0.914	0.926	0.910	0.937	0.904	-	0.933	0.933
NLPR	MAE↓	0.029	0.022	0.025	0.023	0.030	0.028	0.026	0.020	0.023	0.018	0.019	0.024	0.016	0.016
	maxF↑	0.901	0.918	0.903	0.918	0.875	0.908	0.905	0.925	0.923	0.916	0.916	0.883	0.919	0.934
	maxE↑	0.953	0.965	0.956	0.961	0.942	0.952	0.955	0.968	0.965	0.968	0.963	0.954	0.969	0.972
	S↑	0.915	0.931	0.920	0.931	0.898	0.923	0.922	0.936	0.934	0.939	0.922	-	0.942	0.940
SIP	MAE↓	0.054	0.049	0.051	0.055	0.082	0.070	0.060	0.043	0.044	0.034	-	0.047	0.035	0.035
	maxF↑	0.884	0.894	0.879	0.884	0.835	0.857	0.870	0.909	0.901	0.918		0.883	0.916	0.928
	maxE↑	0.920	0.931	0.919	0.922	0.883	0.903	0.909	0.939	0.933	0.944	-	0.926	0.943	0.948
	S↑	0.878	0.885	0.875	0.879	0.816	0.854	0.864	0.896	0.896	0.913	-	-	0.909	0.908
STERE	MAE↓	0.048	0.044	0.039	0.041	0.064	0.045	0.043	0.034	0.038	0.030	0.036		0.030	0.030
	maxF↑	0.881	0.895	0.899	0.903	0.852	0.898	0.897	0.918	0.911	0.902	0.895	-	0.904	0.922
	maxE↑	0.930	0.942	0.944	0.942	0.917	0.942	0.942	0.957	0.949	0.935	0.928	-	0.930	0.958
	S↑	0.892	0.900	0.903	0.908	0.838	0.903	0.901	0.918	0.915	0.925	0.902	-	0.921	0.920

Qualitative experiment

In the **Fig. 4**, this chapter provides a visual comparison that highlights the effectiveness of the DMKAN Net model compared to a series of RGB-D SOD methods. The selected results demonstrate the functionality of the model. In the first two lines, it is particularly evident that there are scenes with low-quality depth cues; Although the depth information is poor, the model in this chapter outperforms other models significantly by successfully depicting prominent objects, which poses challenges for other methods such as BTS Net, CFIDNet, DANet, DMRA, and ICNet, resulting in compromised object segmentation.

The JL-DCF results exhibit diffusion in the salient object boundaries shown in row 3, while AFNet and ICNet only identify a single object of interest in row 4. However, the method proposed in this chapter demonstrates precise delineation of all target objects in such scenarios.

In lines 5 and 6, complex backgrounds are prevalent, and other techniques often struggle to generate precise saliency maps. In contrast, the method proposed in this chapter renders clear and distinct saliency maps while preserving the structural integrity of objects in complex backgrounds. Line 7 demonstrates scenarios with small objects, while line 8 showcases examples with fine-grained objects. Even under these challenging conditions, the method in this chapter maintains robust performance.

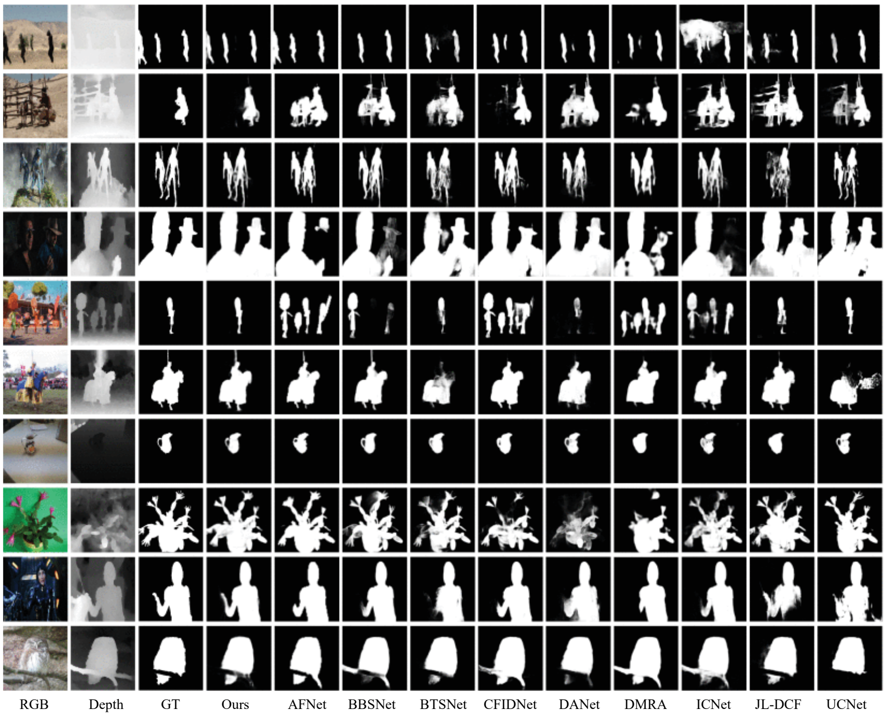

Fig. 4. Visual comparison with state-of-the-art RGB-D models.

Finally, the examples in lines 9 and 10 demonstrate low-contrast scenarios where the foreground and background are highly similar. Almost all competing methods fail to extract the entire object of interest, while the approach proposed in this chapter succeeds. This achievement is attributed to the effective fusion of cross-modal data and the strong fitting capacity of the KAN network, which helps suppress irrelevant background information.

4.3 Ablation Study

Module Ablation

As shown in **Table 2**, we performed an in-depth ablation analysis to verify the effectiveness of each module. DMFAM stands for dual-mode fusion attention module and KAN stands for Tokenized KAN Block. only KAN means removing DMFAM from DMKAN-Net and replacing it with simple concatenation and convolution, only

DMFAM means removing Tokenized KAN Block from DMKAN-Net and replacing it with simple upsampling and convolution operations. origin represents a model that removes both DMFAM and Tokenized KAN Block.

Table 2. Comparison of ablation study results. The best performance in each row is highlighted in **bold**.

Datasets	Metrics	Origin	Only DMFAM	Only KAN	DMKAN-Net
LFSD	MAE ↓	0.095	0.069	0.058	**0.055**
	maxF ↑	0.838	0.847	0.853	**0.885**
	maxE ↑	0.876	0.899	0.908	**0.920**
	S ↑	0.859	0.878	0.897	**0.915**
NJU2K	MAE ↓	0.041	0.031	0.033	**0.025**
	maxF ↑	0.919	0.928	0.930	**0.940**
	maxE ↑	0.944	0.949	0.951	**0.963**
	S ↑	0.913	0.927	0.922	**0.933**
NLPR	MAE ↓	0.029	0.025	0.024	**0.016**
	maxF ↑	0.915	0.927	0.930	**0.934**
	maxE ↑	0.950	0.962	0.964	**0.972**
	S ↑	0.919	0.929	0.932	**0.940**
SIP	MAE ↓	0.047	0.039	0.041	**0.035**
	maxF ↑	0.892	0.916	0.923	**0.928**
	maxE ↑	0.923	0.937	0.935	**0.948**
	S ↑	0.878	0.893	0.891	**0.908**
STERE	MAE ↓	0.041	0.035	0.037	**0.030**
	maxF ↑	0.888	0.911	0.905	**0.922**
	maxE ↑	0.935	0.942	0.943	**0.958**
	S ↑	0.897	0.907	0.913	**0.920**

The original model performed relatively poorly across all datasets, specially on the MAE and S measure. This shows that the original model has some limitations in feature extraction and fusion. After the introduction of DMFAM module, the performance of the model on all data sets is significantly improved. When only the KAN module was used, the model performed better on most datasets than when only the DMFAM module was used. This indicates that KAN module plays an important role in the integration of multi-scale context information and can effectively improve the detection ability of the model. When both DMFAM and KAN modules are used, the model performs best on all datasets. This indicates that the combination of DMFAM and KAN modules can give full play to their respective advantages and further improve the overall performance of the model.

Ablation Study of number of KAN Blocks

In the experiment process of this paper, we directly used 4 KAN blocks to form the KAN Decoder module of the model at the beginning, but at this time, the model training on A40 can only be conducted normally by setting the batch size to 2. Although this setup brings a certain improvement in performance, the time and memory required for training are very expensive. In view of this phenomenon, we conducted ablation experiments on the number of KAN blocks in KAN Decoder to find a good trade-off between Block and reasoning speed. The final results are shown in **Table 3**.

Table 3. Ablation Study of number of KAN Blocks on LFSD . MaxBatchSize means the largest number of trainable Batchsize. Times means the time cost of train phase

Metrics	**1 KANs**	**2 KANs**	**3 KANs**	**4 KANs**
MaxBatchSize ↑	12	8	8	2
Times(h) ↓	14	18	24	58
MAE ↓	0.058	0.056	0.055	0.052
maxF ↑	0.868	0.875	0.885	0.887
maxE ↑	0.904	0.911	0.920	0.922
S ↑	0.888	0.893	0.915	0.918

As can be seen from **Table 3**, when the number of blocks is 3, the accuracy of the model will not be significantly reduced, and the reasoning time will be within the tolerable range. Therefore, we set the number of blocks in the proposed model DMKAN-Net to 3.

5 Conclusion

In this paper, we propose a new dual-modal fusion and KAN Decoder model to solve the RGB-D salient target detection problem. Our framework consists of three parts: a twin Swin Transformer encoder, a dual-modal fusion attention module, and a Tokenized KAN decoder. These three parts have clear division of labor, each performs its own duties, and effectively extracts, merges and fits information. We conducted extensive experiments on our approach on five widely used public datasets. The experimental results show that our method outperforms most of the current state-of-the-art methods on four evaluation indicators. These experimental results fully demonstrate the superiority of our method in handling salient target detection tasks, and its robustness in handling two-modal feature fusion and prediction phases.

However, there is room for further improvement in our approach. For example, although KAN block can fit features well, due to dimensional compression, shallow features will occupy a large amount of video memory when entering KAN block, thus making the training time too long. Our next step is to develop more efficient and memory-friendly KAN modules, and we hope our approach can be applied to other areas as well.

Acknowledgments. This work was funded by the Scientific and technological innovation 2030 major project, China under Grant 2022ZD0115800, and the Xinjiang Uygur Autonomous Region Tianshan Excellence Project, China under Grant 2022TSYCLJ0036.

References

1. Borji, A., Cheng, MM., Hou, Q.: Salient object detection: A survey. Computational Visual Media **5**, 117–150 (2019)
2. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object Detection in 20 Years: A Survey. In: Proceedings of the IEEE, vol. 111, no. 3, pp. 257-276. IEEE (2023)
3. Smeulders, A., Chu, D., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual Tracking: An Experimental Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, **36**(7), 1442-1468 (2014)
4. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep Image Retrieval: Learning Global Representations for Image Search. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science **9910** (2016)
5. Hafiz, A., Bhat, G.: A survey on instance segmentation: state of the art. International Journal of Multimedia Information Retrieval **9**, 171–189 (2020)
6. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y.: Calibrated RGB-D Salient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9471-9481. IEEE (2021)
7. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: RGBD Salient Object Detection via Deep Fusion. IEEE Transactions on Image Processing, **26**(5), 2274-2285 (2017)
8. Wu, Z., Allibert, G., Meriaudeau, F., Ma, C., Demonceaux, C.: HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness. IEEE Transactions on Image Processing **32**, 2160-2173 (2023)
9. Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q.: CIR-Net: Cross-Modality Interaction and Refinement for RGB-D Salient Object Detection. IEEE Transactions on Image Processing **31**, 6800-6815 (2022)
10. Zhang, W., Jiang, Y., Fu, K., Zhao, Q.: BTS-Net: Bi-Directional Transfer-And-Selection Network for RGB-D Salient Object Detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. IEEE, Shenzhen, China (2021)
11. Piao, Y., Ji, W., Yan, G., Li, J., Yao, S., Zhang, M., Cheng, L., Lu, H.: DMRA: Depth-induced Multi-scale Recurrent Attention Network for RGB-D Saliency Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7253-7262. IEEE (2019)
12. Han, J., Chen, H., Liu, N., Yan, C. C., Li, X.: CNNs-Based RGB-D Saliency Detection via Cross-View Transfer and Multiview Fusion. IEEE Transactions on Cybernetics **48**: 3171-3183 (2018)
13. Chen, T., Xiao, J., Hu, X., Zhang, G., Wang, S.: Adaptive fusion network for RGB-D salient object detection. Neurocomputing **522**, 152–164 (2023)
14. Hu, X., Sun, F., Sun, J., Wang, F., Li, H.: Cross-Modal Fusion and Progressive Decoding Network for RGB-D Salient Object Detection. International Journal of Computer Vision **132**, 3067-3085 (2024)
15. Kolmogorov, A.: On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. American Mathematical Society Translations **28**, 55–59 (1957)
16. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T., Tegmark, M.: KAN: Kolmogorov-Arnold Networks. arxiv e-print, arxiv:2404.19756 (2024)
17. Huang, G., Zhao, L., Xing, Y.: Towards theory of deep learning on graphs: Optimization landscape and train ability of Kolmogorov-Arnold representation. Neurocomputing **251**, 10–21 (2017)

18. Liang, X., Zhao, L., Huang, G.: Deep Kolmogorov-Arnold representation for learning dynamics. IEEE Access **6**, 49436–49446 (2018)
19. Xing, Y., Zhao, L., Huang, G.: Kolmogorov-Arnold representation based deep learning for time series forecasting. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1483–1490. IEEE (2018)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012-10022. IEEE (2021)
21. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., Yuan, Y.: U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation, arxiv e-print, arxiv:2406.02918 (2024)
22. Boer, P. D., Kroese, D., Mannor, S., Rubinstein, R.: A Tutorial on the Cross-Entropy Method. Annals Of Operations Research **134**, 19–67 (2005)
23. Mattyus, G., Luo, W., Urtasun, R.: DeepRoadMapper: Extracting Road Topology From Aerial Images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3438-3446. IEEE (2017)
24. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: A benchmark and algorithms. In: European Conference on Computer Vision, pp. 92-109. Springer, Cham (2014)
25. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: IEEE International Conference on Image Processing, pp. 1115-1119. IEEE, Paris, France (2014)
26. Fan, D., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.: Rethinking RGB-D salient object detection: Models data sets and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems **32**(5), pp. 2075-2089 (2021)
27. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2806-2813. IEEE (2014)
28. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 454-461. IEEE, Providence, RI, USA (2012)
29. Ji, W., Li, J., Zhang, M., Piao, Y., Lu, H.: Accurate rgb-d salient object detection via collaborative learning. In: European conference on computer vision, pp. 52–69. Springer, Cham (2020)
30. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time rgb-d salient object detection. In: European conference on computer vision, pp. 646–662. Springer, Cham (2020)
31. Fan, D., Gong, C., Cao, Y., Ren, B., Cheng, M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, Stockholm, Sweden, pp. 698–704. (2018)
32. Fan, D., Cheng, M., Liu, Y., Li, T., Borji, A.: Structuremeasure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp. 4548–4557. IEEE (2017)
33. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE conference on computer vision and pattern recognition, pp. 1597–1604. IEEE, Miami, FL, USA (2009)
34. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE conference on computer vision and pattern recognition, pp. 733–740. IEEE, Providence, RI, USA (2012)

35. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International conference on learning representations. San Diego, CA, USA (2015)
36. Chen, T., Hu, X., Xiao, J., Zhang, G., Wang, S.: CFIDNet: Cascaded feature interaction decoder for RGB-D salient object detection. Neural Computing and Applications **34**(10), 7547-7563 (2022)
37. Li, G., Liu, Z., Ling, H.: ICNet: Information conversion network for RGB-D based salient object detection. IEEE Transactions on Image Processing **29**, 4873-4884 (2020)
38. Zhang, J., Fan, D., Dai, Y., Anwar, S., Saleh, F. S., Zhang, T.: UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8579-8588. IEEE (2020)
39. Fu, K., Fan, D., Ji, G., Zhao, Q., Shen, J., Zhu, C.: Siamese network for RGB-D salient object detection and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5541-5559 (2022)
40. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time RGB-D salient object detection. In: European Conference on Computer Vision, pp. 646-662. Springer, Cham (2020)
41. Fan, D., Zhai, Y., Borji, A., Yang, J., Shao, L.: BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. Proc. In: European Conference on Computer Vision, pp. 275-292. Springer, Cham (2020)
42. Zhang, W., Jiang, Y., Fu, K., Zhao, Q.: BTS-Net: Bi-directional transfer-and-selection network for RGB-D salient object detection. In: International Congress on Mathematical Education (ICME), pp. 1-6. IEEE, Shenzhen, China (2021)
43. Sun, F., Ren, P., Yin, B., Wang, F., Li, H.: CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. IEEE Transactions on Multimedia **26**, 1-14 (2023)
44. Wang, R., Wang, F., Su, Y., Sun, J., Sun, F., Li, H.: Attention-guided multi-modality interaction network for RGB-D salient object detection. ACM Transactions on Multimedia Computing Communications and Applications **20**(3), 1-22 (2023)
45. Wang, F., Wang, R., Sun, F.: DCMNet: Discriminant and cross-modality network for RGB-D salient object detection. Expert Systems With Applications **214** (2023)
46. Tang, B., Liu, Z., Tan, Y., He, Q.: HRTransNet: HRFormer-driven two-modality salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(2), 728-742 (2023)