MutualFormer: Multi-Modality Representation Learning via Mutual Transformer

Xixi Wang, Bo Jiang, Xiao Wang, Jin Tang, and Bin Luo

Abstract—Aggregating multi-modality data to obtain accurate and reliable data representation attracts more and more attention. The pristine researchers generally adopt the CNN to extract features of independent modality and aggregate them with a fusion module. However, the overall performance is becoming saturated due to limited local convolutional features. Recent studies demonstrate that Transformer models usually work comparable or even better than CNN for multi-modality task, but they simply adopt concatenation or cross-attention for feature fusion which may just obtain sub-optimal results. In this work, we re-thinking the self-attention based Transformer and propose a novel MutualFormer for multi-modality data fusion and representation. The core of MutualFormer is the design of both token mixer and modality mixer to conduct the communication among both tokens and modalities. Specifically, it contains three main modules, i.e., i) Self-attention (SA) for intra-modality token mixer, ii) Cross-diffusion attention (CDA) for inter-modality mixer and iii) Aggregation module. The main advantage of the proposed CDA is that it is defined based on individual domain similarities in the metric space which thus can naturally avoid the issue of domain/modality gap in cross-modality similarities computation. We successfully apply the MutualFormer into the saliency detection problem and propose a novel approach to obtain the reinforced features of RGB and Depth images. Extensive experiments on six popular datasets demonstrate that our model achieves comparable results with 16 SOTA models.

Index Terms—Multi-Modality, Transformer, Self-attention, Cross-diffusion Attention, RGB-D Salient Object Detection.

1 INTRODUCTION

As we all know, a single modality may only work well in certain situations and express monotonous and limited information. For example, the thermal camera only perceives the temperature of the object’s surface; the depth camera reflects the spatial distance information; the RGB camera captures the color and texture information, but performs poor under low-illumination and fast motion, etc. Therefore, how to fuse information from various modalities and achieve more reliable performance draws more and more attention in recent years.

The fundamental challenge of multi-modality fusion is how to exploit the useful cues of both intra-modality and cross-modality simultaneously for final representation. According to our observation, existing works mainly focus on proposing various fusion methods including early [1], [2], [3], middle [4], [5], [6], and late fusion [7], [8]. Among them, the middle-level (or feature-level) is the most popular fusion strategy due to its good balance of flexibility and accuracy. These methods usually first utilize several separate backbone networks for various modalities to conduct intra-modality feature learning. Then, an information fusion module is designed for learning and aggregating cross-modality features. After reviewing the early deep learning algorithms on multi-modality data representation, we can find that they usually adopt CNN as the backbone, which learns the local features well with the help of convolutional operators. However, their performance is becoming saturated due to the limited representation of local CNN features [9], [10].

Recently, Transformer has shown its strong ability in modeling the dependence of tokens (e.g., image patches) based on the token mixer architecture which is usually achieved via a self-attention module [11]. It is firstly proposed for natural language processing [12] in a self-supervised manner. Then, the Transformer based Computer Vision (CV) models [13], [14], [15], [16], [17], [18] also obtain the top-k ranks on many benchmarks and tasks. More and more works demonstrate that Transformer can achieve comparable or even better performance than CNN. In addition to single modality based problems, researchers also introduce the Transformers into the multi-modality tasks [19], [20], [21], [22], [23], [24]. However, existing works adopt simple concatenation [20], [22], [24], or intuitive cross-modality fusion [19], [21], [23], [25], which may obtain sub-optimal results. Because the first strategy generally fails to fully exploit the relationship between the two modalities; the second way directly connects various modalities with simple dot product, which may result in unreliable learning results caused by the modality gap. Therefore, Transformer-based multi-modality fusion is still an interesting problem to be further studied.

In this paper, we propose a novel Transformer based multi-modality data representation method, termed MutualFormer. The core of MutualFormer is our new strategy to conduct the communication between tokens and modal-
In this section, we give a brief review on the Multi-modality Transformer and RGB-D salient object detection.

**Multi-modality Transformer.** As the self-attention based Transformer network models the global long-range dependence, it can achieve comparable or better performance than CNN. Many researchers attempt to introduce the Transformer into multi-modality problems \cite{19, 20, 21, 22, 24, 28, 29, 30, 31, 32}. Part of them adopt BERT-like \cite{33} framework which concatenates multimodalities along the channel dimension \cite{24, 34} or sums the tokens of each modality \cite{20} before input. Although good performance can be obtained, however, these methods only conduct the self-similarity computation to obtain the coarse fused features which may limit their final results.

Some works also attempt to handle this problem with cross-attention mechanism \cite{23, 25, 35}. For example, Tan \textit{et al.} \cite{19} propose LXMERT framework for learning the connections of vision and language semantics for vision-and-language reasoning. Yu \textit{et al.} \cite{36} present a Multi-modal Transformer (MT) framework by using an image encoder and a caption decoder for image captioning. Nagrani \textit{et al.} \cite{22} propose Multimodal Bottleneck Transformer (MBT) used latent units to fuse the cross-modal information for audiovisual fusion. Feng \textit{et al.} \cite{37} propose a multimodal transformer (MTrans) that leverages a modified attention mechanism, i.e., cross attention, to fuse the complementary information of multi-modal for accelerated magnetic resonance imaging. Curto \textit{et al.} \cite{38} develop a multi-modal multi-subject Transformer architecture (DyadFormer) to fuse both video and audio modality for dyadic interactions. Gabeur \textit{et al.} \cite{28} introduce a Multi-modal Transformer (MMT) to jointly encode the representation of vision and caption in the video for video retrieval. Liu...
et al. [27] propose a MultiModal Transformer (mMTransformer) framework based on stacked transformers to aggregate multiple channels of contextual information for multi-modal motion prediction.

Compared with concatenate based multi-modal Transformers, the cross-attention based methods can model intra- and inter-modality to obtain better feature representations. Therefore, they can achieve better performance than the concatenate based algorithms. However, the cross-modality similarities computation is generally defined by the feature representations of multi-modality token, seldom of them considers the domain/modality gap between features of different domains. Thus, these models maybe obtain sub-optimal results only.

RGB-D SOD. RGB-D salient object detection aims to locate the most salient objects (or regions) from visual image(s). Recent SOD methods adopt Transformer or Hybrid CNN-Transformer to achieve better results, as mentioned in previous sub-sections. Before that, researchers tend to utilize CNN for feature learning and fusion. For example, Pang et al. [38] propose a hierarchical dynamic filtering network (HDFNet) which mainly respectively employ dynamic dilated pyramid module to generate the adaptive kernel in multi-level for decoding RGB feature. Li et al. [39] exploit a Hierarchical Alternate Interaction Module (HAIM), which follows the RGB-deep-RGB flow for optimizing depth features based on RGB features guides. For RGB features, it adopts a hierarchical way to enhance. Luo et al. [7] design Cascade Graph Neural Networks (Cas-GNN) based on Graph techniques to model the relationships between multi-modal information. Piao et al. [8] introduce a depth distiller (A2dele) to transfer depth knowledge and the localization knowledge of salient objects from the depth stream to RGB stream. Fu et al. [40] propose joint learning and densely-cooperative fusion framework based on Siamese network of sharing parameters. Ji et al. [11] design a collaborative learning framework (CoNet) which leverages edge, depth and saliency in mutual-benefit learning manners to help generate accurate saliency results. Liu et al. [6] propose a Selective Self-Mutual Attention to select reliable information of each modality for fusing learning. At the current stage, these methods are outperformed by Transformer based SOD detectors, as the limited representation ability of local convolutional filters. In this paper, our proposed RGB-D SOD framework jointly learns the local-global features with FFE based CNN encoder and MutualFormer based Transformers. Therefore, we can get more robust and accurate features for SOD, which bring us more accurate SOD results.

3 THE PROPOSED MUTUALFORMER

As shown in Fig. 1, Transformer has strong ability in modeling the dependence of tokens (for example, image patches, video frames, etc) based on the token mixer architecture which is usually achieved via a self-attention (SA) module [13], [15], [33]. In this section, we extend this mechanism to Mutual Transformer (MutualFormer) to further model the dependence of multiple modalities for multi-modality data representation problem. The core of MutualFormer is the design of both inter-modality token mixer and cross-modality token mixer to conduct the information communication between tokens and modalities simultaneously. The whole architecture of MutualFormer is shown in Fig. 1 which contains 1) Self-attention (SA) for intra-modality token mixer, 2) Cross-diffusion Attention (CDA) for inter-modality token mixer and 3) Aggregation for final token representation.

3.1 Self-attention (SA) for intra-modality token mixer

Let $X_r \in \mathbb{R}^{n \times d}$ and $X_d \in \mathbb{R}^{n \times d}$ represent the token set of two modalities respectively, where $n$ denotes the number of tokens in both modalities and $d$ is the feature dimension of tokens. To capture the dependence of tokens within each modality, we utilize the commonly used self-attention architecture [12], [13] for each modality. Specifically, we first compute the similarities among different tokens for each modality as

$$S_r = \text{Softmax}(Q_rK_r^T)$$

$$S_d = \text{Softmax}(Q_dK_d^T)$$

where $Q_r$ and $K_r$ are obtained by conducting two linear transformations on $X_r$ respectively. Similar to the acquisitions $Q_d$ and $K_d$ in Eq. (2). Based on $S_r$ and $S_d$, we then can obtain context-aware representations $M_r$ and $M_d$ for tokens in each modality by using the self-attention (SA) mechanism as

$$M_r = S_rV_r, \quad M_d = S_dV_d$$

where $V_r$ and $V_d$ are obtained by conducting linear transformations on $X_r$ and $X_d$ respectively.

3.2 Cross-diffusion Attention (CDA) for inter-modality token mixer

The key aspect of the proposed MutualFormer is to conduct the information communication among tokens belonging to different modalities. One popular way to address this issue is to design some cross-attention mechanisms [19], [21], [23], [25], [42] and the core of cross-attention is the computation of cross-similarities $S_{rd}$ and $S_{dr}$ among tokens of different modalities. One simple way to compute $S_{rd}$ and $S_{dr}$ is $S_{rd} = S_{dr} = S_r + S_d$ as discussed in work [6]. In addition, some works [19], [21] also propose to define $S_{rd}, S_{dr}$ based on features of tokens and define Cross-attention (CA) model as,

$$S_{rd} = \text{Softmax}(Q_rK_d^T), \quad M_{rd} = S_{rd}V_d$$

$$S_{dr} = \text{Softmax}(Q_dK_r^T), \quad M_{dr} = S_{dr}V_r$$

However, since $Q_r$ and $K_r$ are derived from different modalities, therefore, the intrinsic modality/domain gap makes $Q_rK_d^T$ be unreliable to reflect the pairwise similarities between different domains.

To overcome this issue, inspired by Regularized Diffusion Process (RDP) [26], we propose to define a novel Cross-diffusion Attention (CDA). Instead of defining cross similarities on feature space, CDA is defined on metric space, as shown in Fig. 2. Specifically, let $\hat{S}_r = D_r^{-\frac{1}{2}}S_rD_r^{-\frac{1}{2}}$ and $\hat{S}_d = D_d^{-\frac{1}{2}}S_dD_d^{-\frac{1}{2}}$ denote the normalized similarity matrices where $D_r, D_d$ is a diagonal matrix with elements

$$D_r = \text{Diag}(s_{r1}, ..., s_{rn})$$

$$D_d = \text{Diag}(s_{d1}, ..., s_{dn})$$
defined by the row-addition of \( S_r, S_d \). We propose to define our novel Cross-diffusion Attention (CDA) as

\[
S_{rd} = \epsilon S_r \hat{S}_r^{(0)} \hat{S}_d^T + (1 - \epsilon) A, \quad M_{rd} = S_{rd} V_d \tag{6}
\]
\[
S_{dr} = \epsilon S_d \hat{S}_d^{(0)} \hat{S}_r^T + (1 - \epsilon) A, \quad M_{dr} = S_{dr} V_r \tag{7}
\]

where \( A = \frac{1}{2}(S_r + S_d) \) and parameter \( \epsilon \in (0, 1) \) denotes the balancing hyper-parameter. \( S_{rd}^{(0)} \) and \( S_{dr}^{(0)} \) can be initialized as the identity matrix \( I \) or some other initial affinity matrices obtained by using other approaches, e.g., affinity matrix \( A \). In this paper, we simply set \( S_{rd}^{(0)} \) and \( S_{dr}^{(0)} \) to \( I \). In this case, Eqs. (6, 7) is simply defined as

\[
S_{rd} = \epsilon S_r \hat{S}_r T + (1 - \epsilon) A, \quad M_{rd} = S_{rd} V_d \tag{8}
\]
\[
S_{dr} = \epsilon S_d \hat{S}_d T + (1 - \epsilon) A, \quad M_{dr} = S_{dr} V_r \tag{9}
\]

where \( V_r \) and \( V_d \) are respectively obtained by conducting linear transformations on \( X_r \) and \( X_d \).

Remark. The main advantages of the proposed CDA model Eqs. (6, 7) are described as below:

- **Reliable.** The proposed CDA is defined based on individual domain similarity \( S_r \) and \( S_d \) in metric space which thus can naturally avoid the issue of domain/modality gap in cross-modality similarities computation. Comparing with feature-based cross-attention (CA) methods, such as Eqs. (4, 5) [19], [21], [25], the proposed CDA is more reliable.

- **Flexible.** The above CDA provides a general scheme. In this paper, we simply set \( A = \frac{1}{2}(S_r + S_d) \) and \( S_r^{(0)} = S_d^{(0)} = I \). In real applications, one can derive many specific CDA models by initializing \( A \) and \( S_r^{(0)}, S_d^{(0)} \) with different approaches. That is, CDA can also serves as a post-processing procedure for other CA methods (such as Eqs. (4, 5)) to learn more reliable cross-modality attentions.

- **Efficiency.** In Transformer, since token number \( n \) is usually less than feature dimension \( d \), thus CDA Eqs. (6, 7) is usually implemented more efficiently than traditional feature based CA method Eqs. (4, 5).

After using the above SA and CDA modules, we can obtain the representations \{\( M_r, M_{rd} \)\} and \{\( M_d, M_{dr} \)\} for Modal-\( r \) and Modal-\( d \) which encode the spatial and modality-context information respectively. By integrating \( M_r \) and \( M_{rd} \) together, we can obtain more reliable representations \( H_r \) of tokens in Modal-\( r \) by incorporating both spatial and modality-context information simultaneously, i.e.,

\[
H_r = f_r(M_r \parallel M_{rd}) \tag{10}
\]

We can similarly obtain \( H_d \) in Modal-\( d \) as

\[
H_d = f_d(M_d \parallel M_{dr}) \tag{11}
\]

where \( \parallel \) denotes the concatenation operation. \( f_r(\cdot) \) and \( f_d(\cdot) \) represent the two convolutional layers with different parameters.

### 3.3 Aggregation block

The final output of MutualFormer is the modality-invariant and context-aware representations \( P \) for tokens. We obtain \( P \) by aggregating \( H_r \) and \( H_d \) together as

\[
P = FFN \left( LN \left( g(H_r \parallel H_d) \right) \right) + h(X_r \parallel X_d) \tag{12}
\]

where \( g(\cdot) \) and \( h(\cdot) \) denote the two convolutional layers and \( \parallel \) denotes the concatenation operation. FFN denotes the two fully-connected layers with a non-linearity activation function GELU. LN represents the layer-normalization operation.

### 4 RGB-D Salient Object Detection

To validate the effectiveness of our newly proposed MutualFormer, we select the RGB-Depth salient object detection which is a popular problem. As shown in Fig. 3, the MutualFormer-based SOD framework contains three main modules, including Focal Feature Extractor (FFE), MutualFormer-based Fusion module (MF) and Decoder network. We will introduce these modules in subsequent sections, respectively.

#### 4.1 Focal Feature Extractor

Given the input RGB \( X_r \) and Depth \( X_d \) image pairs, we need to first extract their multi-level feature representations, respectively. Without loss of generality, we adopt the widely used ResNet-50 [43] as the backbone network. The parameters are initialized with pre-trained models on ImageNet [44] classification task. Therefore, we can get the multiscale feature maps \( F_r^l \in \mathbb{R}^{H^l \times W^l \times C^l}, F_d^l \in \mathbb{R}^{H^l \times W^l \times C^l}, l \in \)
Fig. 3. Architecture of the proposed network for RGB-D salient object detection. The whole network consists of Focal Feature Extractor (FFE), MutualFormer-based Fusion (MF) module and Decoder. Specifically, FFE uses ResNet-50 [33] pre-trained on ImageNet [44] as the backbone with the supervision of focal regularization to obtain enhanced CNN representation. $F^3_r$ and $F^3_d$ are the outputs of different levels in two branches. The MF is our proposed fusion module, which can be seen more details in Section 3, therefore, we only describe the network architectures and implementation details in the decoder network as described in Eq. (12) by implementing which consists of weighted binary cross-entropy.

\begin{equation}
\mathcal{L}_{R,r,d} = - \frac{1}{M} \sum_{i=1}^{M} \alpha_i \left(1 - p^i_r \log \left( \frac{p^i_r}{p^i_r + p^i_d} \right) \right) \log(\hat{p}^i_r) \tag{13}
\end{equation}

where $M = H \times W$ denotes the number of pixels. $p^i_r, p^i_d \in [0, 1]$ represent the foreground probability of the $i$-th pixel which is obtained based on $F^l_r$ and $F^l_d$ respectively. The weighting factor $\alpha_i$ is set as 0.25 for foreground and 0.75 otherwise in our experiments. Similarly, we can also obtain the regularization loss $\mathcal{L}_D$ for Depth branch.

### 4.2 MutualFormer-based Fusion

As illustrated in Fig. 3, we apply our newly proposed MutualFormer for the feature fusion in RGB-Depth SOD. Due to we have introduced the motivation and working principle of MutualFormer in Section 3 therefore, we only describe the network architectures and implementation details in the following paragraphs.

In order to achieve intra-modality enhancement and cross-modality interaction, we introduce the proposed MutualFormer module. Given the RGB-Depth features, $F_r$ and $F_d$, we first calculate the intra-modality features with self-attention layers for each modality. Specifically, we embed the input features $F_r, F_d$ into $\tilde{F}_r, \tilde{F}_d$ via convolutional operations. Also, we add the position encoding $P_r$ and $P_d$ with its corresponding feature embedding. Then, the multi-head self-attention layers are adopted for intra-modality feature learning and thus can respectively obtain the intra-modality features $M_r$ and $M_d$ for RGB-modality and Depth-modality.

For multi-modality learning tasks, such as RGB-D SOD, it is necessary to conduct feature learning from multiple modalities simultaneously to integrate information complementarity. Specifically, we compute the CDA based on the self-similarities of RGB and Depth modality and thus obtain the enhanced cross-modality representation $M_{rd}$ and $M_{dr}$. Next we can fuse the representation of intra-modality and inter-modality to obtain the output by Eqs. (10, 11), i.e., $H_r$ and $H_d$.

To speed up the fitting speed of training, we also add a single Transformer network. Specifically, we first concatenate the two input features and fed into a convolutional layer to obtain a unified representation. After that, following the mainstream of Transformers [13, 17] to get the coarse fused features $H$. Finally, we obtain the final representation of the whole Encoder network as described in Eq. (12) by integrating the above representations together, followed by residual connections, Layer Normalization (LN) and Feed-Forward Networks (FFN).

### 4.3 Decoder Network

In the decoding phase, following [47], we feed the multi-level fused features into a decoder network to obtain corresponding decoding features. The decoder contains two phases, the first phase takes the input features and output the feature map using Cross-Level Modules (CLM). CLM aims to fuse features of different levels, more details can be found in [47]. Then, the features predicted from first decoding phase will be used to guide the second decoding phase.

### 4.4 Loss Function

We use the pixel position aware loss [47] for saliency prediction which consists of weighted binary cross-entropy.
loss and weighted intersection-over-union loss. Formally, for each \( l \)-level prediction, the pixel position aware loss \( \mathcal{L}_P^l \) is computed as

\[
\mathcal{L}_P^l = \frac{1}{2}(\omega \mathcal{L}_{ce}^l + \mathcal{L}_{wiou}^l)
\]

where \( \mathcal{L}_{ce}^l \) and \( \mathcal{L}_{wiou}^l \) denote the standard binary cross-entropy loss and weighted IOU loss respectively. \( \omega = 1 + 5 |AP(y) - y| \) with \( AP(\cdot) \) performing average pooling operation and \( y \) represents ground truth. As a result, the loss for saliency prediction is formulated as

\[
\mathcal{L}_P = \frac{1}{m} \sum_{i=1}^{m} \sum_{l=2}^{5} \frac{1}{2} \mathcal{L}_P^l
\]

where \( m \) is the total number of sub-decoder, \( i \) and \( l \) denote \( i \)-th sub-decoder and \( l \)-th level respectively. By adding the regularization loss discussed in Eq. (15), the final whole loss of our model is formulated as

\[
\mathcal{L}_{total} = (1 - \lambda) \mathcal{L}_P + \lambda \sum_{l=2}^{5} \left( \mathcal{L}_{R_{c,d}}^l + \mathcal{L}_{R_{d,r}}^l \right)
\]

where \( \lambda \in \{0, 0.9\} \) stands for the balancing parameter, which is set to 0.4 in all our experiments. \( l \) denotes the \( l \)-th level.

5 EXPERIMENTS

5.1 Datasets and Evaluation Metrics

Datasets. We evaluate the proposed RGB-D saliency detection method on six commonly used datasets, including NJU2K [49] (1985 image pairs), SIP [50] (929 image pairs), NLPR [51] (1000 image pairs), LFsD [52] (100 image pairs), STEREO [53] (797 image pairs), DUT-RGBD [1] (1200 image pairs), which are all composed of RGB and Depth image pairs. Following previous works [1], [5], [8], we select 1485, 700 and 800 image pairs from NJU2K [49], NLPR [51] and DUT-RGBD [1] respectively as the training data. The remaining image pairs are used as the test set. We employ horizontal flip and random crop as used in [2], [9] for training data augmentation and thus obtain a final training set that contains 32835 RGB-D image pairs.

Evaluation Metrics. We utilize five commonly used evaluation metrics, i.e., Precision-Recall (PR) curve [54], S-measure \( (S_m) \) [55], maximum F-measure \( (F_{max}) \) [56], maximum E-measure \( (E_{max}) \) [57] and Mean Absolute Error \( (MAE) \) [54], [58], to evaluate the performance of different RGB-D SOD models. Specifically, the Precision and Recall are defined as

\[
\text{Precision} = \frac{TP}{TP + FP}; \quad \text{Recall} = \frac{TP}{TP + FN}
\]

where TP, FP and FN respectively represent true positives, false positives and false negatives. S-measure [55] calculates the structure similarity between the predicted saliency result and ground truth. It jointly considers the similarities of object structure and region structure as,

\[
S\text{-measure} = \alpha S_{object} + (1 - \alpha) S_{region}
\]

where we set \( \alpha = 0.5 \) empirically. F-measure [56] calculates the harmonic mean of average precision and recall values which can be obtained as,

\[
F = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} \times \text{Recall}}
\]

where \( \beta^2 = 0.3 \) by following previous works [50]. E-measure [57] combines image level statistics with local pixel matching information to evaluate the saliency binary map. MAE [58] refers to average pixel-wise absolute error between the saliency results and ground truth results. The evaluation code we used is available at [http://dpfan.net/d3netbenchmark/](http://dpfan.net/d3netbenchmark/).

5.2 Implementation Details

For RGB-D salient object detection, we employ ResNet-50 [43] with fully connected layers removed as CNN feature extractor which is pre-trained on ImageNet [44] classification task. All the image pairs are resized to 352 × 352 and then input to the corresponding backbone to obtain its features. We adopt Adam [59] optimizer to train our model. The initial learning rate is set to 5e-6 for backbone and 4.5e-5 for others, which is reduced by multiplying 0.1 after 5-th epoch and 0.2 after 8-th epoch respectively. For our fusion module, we borrow ViT [13] model trained on ImageNet [44] to initialize some parameters of MutualFormer. We experimentally set patch size as (11, 11) for all levels and the number of patches \( N_l \) as \{100, 81, 36, 1\} when \( l = \{2, 3, 4, 5\} \). In our model, the channel \( c \) of each patch is set to 64. The batch size is set to 11 and train the whole network with 15 epochs. In the testing phase, both RGB and Depth input image pairs are first resized to 352 × 352 resolution and then feed into the proposed network to output the corresponding saliency map. The proposed network is implemented by PyTorch 1.7 on a single 11G NVIDIA RTX2080Ti GPU. Our code will be released at [https://github.com/SissiW/MutualFormer](https://github.com/SissiW/MutualFormer).

5.3 Comparison with State-of-the-Art Methods

We compare our proposed network with some other recent state-of-the-art methods, i.e., CoNet [41], cmMS [4], D3Net [50], HDFNet [58], FRDT [3], CasGnn [7], S3F [5], A2dele [8], S2MA [6], HAINet [39], DQSD [60], RD3D [10], DCF [61], CDNet [62], JLDCF [40] and TriTrans [20].

Quantitative Results. For fair comparisons, all the results are evaluated by using the same evaluation code provided in [50]. Table I shows the quantitative results of four metrics on all datasets. We can observe that our proposed network generally achieves better performance than other recent RGB-D SOD methods on NJU2K [49], NLPR [51], DUT-RGBD [1], SIP [50], STEREO [53] and LFsD [52] datasets. This demonstrates the effectiveness of the proposed RGB-D saliency detection network. Meanwhile, Fig. 4 shows the comparisons on Precision-Recall (PR) curve, which demonstrates that the proposed network method generally achieves better performance than other related methods on all six benchmark datasets.

Qualitative Results. Fig. 5 shows some saliency map examples of ours and the compared methods. We can find

1. These methods have publicly released codes or results.
that the proposed network obtains better detection results in the various challenging scenarios, including complex backgrounds ($1^{st} - 2^{nd}$), salient object and background with similar appearance ($3^{rd} - 4^{th}$), multiple objects ($5^{th} - 6^{th}$), low-quality depth map ($7^{th} - 8^{th}$) and small object ($9^{th} - 10^{th}$). It further demonstrates that our proposed method
| # | T | NJU2K | NLPR | DUT-RGBD | LFSD |
|---|---|---|---|---|---|
| 0 | T = 1 | 0.923 | 0.925 | 0.956 | 0.032 |
| 1 | T = 2 | 0.922 | 0.923 | 0.954 | 0.032 |
| 2 | T = 3 | 0.921 | 0.923 | 0.953 | 0.032 |
| 3 | T = 4 | 0.920 | 0.921 | 0.950 | 0.034 |
| 4 | T = 5 | 0.921 | 0.922 | 0.954 | 0.033 |

**TABLE 2**
Ablation study for MutualFormer Fusion layers in NJU2K, NLPR, DUT-RGBD and LFSD datasets. The best results are bolded.

**TABLE 3**
Ablation study for different fusion strategies in four datasets. The “Add/Cat” denotes directly using a simple adding/concatenating operation to fuse the multi-modality features. The “CrossFormer” represents the self-attention mechanism of the standard Transformer replaced by cross-attention mechanism of Eqs. (4-5) for multi-modality fusion. The best results are bolded.

| Fusion Strategy | NJU2K | NLPR | DUT-RGBD | LFSD |
|---|---|---|---|---|
| Add | 0.940 | 0.907 | 0.940 | 0.038 |
| Cat | 0.908 | 0.902 | 0.938 | 0.039 |
| Transformer | 0.917 | 0.917 | 0.949 | 0.034 |
| CrossFormer | 0.918 | 0.919 | 0.952 | 0.035 |
| CrossFormer_CDA | 0.918 | 0.920 | 0.952 | 0.033 |
| MutualFormer | 0.922 | 0.923 | 0.954 | 0.032 |

**TABLE 4**
Ablation study for the different components of proposed network. The “MF” is the proposed MutualFormer fusion strategy. The “FFE” denotes the proposed focal feature extractor. The best results are bolded.

| # | Baseline | MF | FFE | NJU2K | NLPR | DUT-RGBD | LFSD |
|---|---|---|---|---|---|---|---|
| 1 | ✓ | ✓ | ✓ | 0.916 | 0.901 | 0.935 | 0.041 |
| 2 | ✓ | ✓ | ✓ | 0.910 | 0.907 | 0.940 | 0.038 |
| 3 | ✓ | ✓ | ✓ | 0.920 | 0.921 | 0.952 | 0.033 |
| 4 | ✓ | ✓ | ✓ | 0.922 | 0.923 | 0.954 | 0.032 |

**TABLE 5**
Comparison results of different parameter $\lambda$ in Eq. (16).

| $\lambda$ | NJU2K | NLPR | DUT-RGBD | LFSD |
|---|---|---|---|---|
| 0.0 | 0.920 | 0.921 | 0.952 | 0.033 |
| 0.1 | 0.918 | 0.920 | 0.950 | 0.033 |
| 0.2 | 0.917 | 0.919 | 0.950 | 0.034 |
| 0.3 | 0.917 | 0.917 | 0.951 | 0.034 |
| 0.4 | 0.922 | 0.923 | 0.954 | 0.032 |
| 0.5 | 0.917 | 0.918 | 0.949 | 0.034 |
| 0.6 | 0.919 | 0.921 | 0.950 | 0.034 |
| 0.7 | 0.920 | 0.921 | 0.950 | 0.033 |
| 0.8 | 0.917 | 0.918 | 0.947 | 0.035 |
| 0.9 | 0.918 | 0.920 | 0.950 | 0.035 |

**TABLE 6**
Comparison results for different parameter $\epsilon$ in Eqs. (6-7).

| $\epsilon$ | NJU2K | NLPR | DUT-RGBD | LFSD |
|---|---|---|---|---|
| 0.0 | 0.915 | 0.915 | 0.949 | 0.035 |
| 0.1 | 0.918 | 0.920 | 0.950 | 0.033 |
| 0.2 | 0.917 | 0.919 | 0.950 | 0.034 |
| 0.3 | 0.917 | 0.914 | 0.947 | 0.035 |
| 0.4 | 0.915 | 0.913 | 0.946 | 0.035 |
| 0.5 | 0.917 | 0.920 | 0.949 | 0.034 |
| 0.6 | 0.922 | 0.923 | 0.954 | 0.032 |
| 0.7 | 0.915 | 0.914 | 0.947 | 0.035 |
| 0.8 | 0.920 | 0.922 | 0.952 | 0.033 |
| 0.9 | 0.917 | 0.918 | 0.950 | 0.034 |
| 1.0 | 0.918 | 0.920 | 0.950 | 0.034 |
can well capture both intra-modality specific representation and cross-modality complementary information for RGB-D based SOD task.

5.4 Ablation Study
In this section, we conduct extensive experiments on four datasets (NJU2K, NLPR, DUT-RGBD and LFSD) to help
Impact of Multi-layer MutualFormer. To verify the effect of multi-layer MutualFormer in each level for RGB-D SOD task, we set different layer numbers $T$ to conduct ablation study. As shown in Table 2, we list the comparison results of $T$ gradually increasing from 1 to 5. We can see that the performance of $T = 2$ generally outperforms $T = 1$ except for NJU2K, which demonstrates the more effectiveness of the multi-layer MutualFormer architecture in each level. However, the performance decreases when the layer $T$ of MutualFormer is larger than 2. As a result, we set the number of MutualFormer layers to 2 for each level in our experiments.

Analysis of Different Fusion Strategies. In order to analyze the effectiveness of different fusion strategies (i.e., Add, Concatenate (Cat), Transformer, CrossFormer, CrossFormer$_{CDA}$ and MutualFormer), the entire network architecture remains unchanged except for the fusion module. The Add, Cat, Transformer and CrossFormer are the four commonly used multi-modality fusion strategies, where Add/Cat means directly using a simple adding/concatenating operation to fuse the two modality features. The Transformer represents we employ the standard Transformer [12] as the fusion module. The CrossFormer denotes the self-attention mechanism of the standard Transformer replaced by the cross-attention mechanism (CA) of Eqs. (4, 5) for multi-modality fusion. The CrossFormer$_{CDA}$ employs the cross-diffusion attention (CDA) in Eqs. (6, 7). All the results are shown in Table 5. We can observe that: 1) All the transformer-based fusion methods obtain surprising results, which indicates the effectiveness of transformer architecture for the fusion multi-modality data. 2) Comparing CrossFormer$_{CDA}$ with CrossFormer, we observe that CrossFormer$_{CDA}$ can obtain better performance. For example, the CrossFormer$_{CDA}$ obtain 0.923 on $F_{max}$, meanwhile, the CrossFormer achieves 0.917 on this metric based on NLPR dataset. This proves the effectiveness of our proposed CDA. 3) The performance of the proposed MutualFormer is better than other methods based on transformer architecture. This demonstrates the effectiveness of the proposed MutualFormer architecture.

Impact of Different Components. To validate the effectiveness of each component in the proposed network, we implement some variants of the proposed model as follow: (1) Baseline represents using a simple adding operation to fuse the two modality features. (2) MF denotes using our proposed MutualFormer fusion module to replace the fusion module of Baseline for fuse multi-modality features. (3) FFE refers to the standard feature extractor adding the supervision of pixel-level focal regularization.

All the experimental results of these variants are provided in Table 4. Specifically, we can observe that: (1) Comparing the results of # 1 and # 2 (# 3 and # 4), we can see that FFE performs better than without it. It demonstrates the effectiveness of the proposed focal feature extractor. (2) The performance of our proposed MF is better than Baseline. Our performance can be further improved by introducing the FFE, which fully verified the effectiveness of each component in our architecture.

5.5 Parameter Analysis

In our proposed RGB-D SOD framework, two parameters are important for the final performance, i.e., the $\epsilon$ in Eqs. (6, 7), and the $\lambda$ in Eq. (16). In this section, we test our model with various values of the two parameters to check their influence. Specifically, we set the $\epsilon \in \{0, 1\}$ and $\lambda \in \{0, 0.9\}$. As shown in Table 5 and Table 6, we can find that when the $\lambda$ and $\epsilon$ are set as 0.4 and 0.6, our model achieves better results on the four datasets, including NJU2K, NLPR, DUT-RGBD, and LFSD. To give a more intuitive presentation, we also visualize these results in Fig. 5 and Fig. 6.
5.6 Feature Visualization

According to aforementioned experimental analysis on multiple benchmark datasets, we can find that our proposed MutualFormer indeed helps the RGB-D SOD task. To help readers have a more intuitive understanding of our model, we give some visualizations of feature maps of our MutualFormer and compared baselines, including Add, Concatenate (Cat), Transformer, CrossFormer, and CrossFormer_CDA. As shown in Fig. 8, we can find that our proposed MutualFormer can obtain better visual effects, which further validated the effectiveness and advantages of our model.

5.7 Limitation Analysis

Although our proposed model achieves better results on multiple datasets as validated in previous sections, however, our model also performs poorly in some very challenging cases. As shown in Fig. 9, our model error detecting the footstool at the upper row, and fail to segment the fine-grained lines in the bicycle wheels. These cases are really challenging for current SOD detectors, and other state-of-the-art methods also perform badly. In our future works, we will consider borrowing fine-grained detection algorithms to further improve final results.

![RGB Depth GT HAINet RD3D DCF TriTrans OURS](image)

Fig. 9. Bad cases of our RGB-D SOD algorithms.

6 Conclusion

In this paper, we propose a novel Mutual Transformer (MutualFormer) to conduct the interaction between tokens and modalities for multi-modality task. MutualFormer mainly contains three modules: Self-attention (SA), Cross-diffusion Attention (CDA) and Aggregation module. Specifically, MutualFormer first computes the self-similarity of single-modality and then performs the self-attention mechanism for modeling the relationship of intra-modality tokens. It simultaneously performs the cross-diffusion attention to model the relationship of inter-modality tokens based on individually modal similarity in metric space, which effectively avoids the issue of domain/modality gap. Finally, we leverage an aggregation module to obtain the final enhanced feature representations. In order to validate the effectiveness of our proposed MutualFormer, we take the RGB-D salient object detection as an example to conduct extensive experiments. In addition, we introduce a Focal Feature Extractor (FFE) to obtain the reinforced features of RGB and Depth image for better fusion. Extensive experimental results on six datasets prove the effectiveness of our proposed model.

References

[1] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, “Depth-induced multi-scale recurrent attention network for saliency detection,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7254–7263.

[2] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, “Contrast prior and fluid pyramid integration for rgbd salient object detection,” in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2019, pp. 3922–3931.

[3] M. Zhang, Y. Zhang, Y. Piao, B. Hu, and H. Lu, “Feature reintegration over differential treatment: A top-down and adaptive fusion network for rgb-d salient object detection,” in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4107–4115.

[4] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, “Rgb-d salient object detection with cross-modality modulation and selection,” in Proceedings of European Conference on Computer Vision, 2020, pp. 225–241.

[5] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, “Select, supplement and focus for rgb-d saliency detection,” in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2020, pp. 3472–3481.

[6] N. Liu, N. Zhang, and J. Han, “Learning selective self-mutual attention for rgb-d saliency detection,” in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2020, pp. 13753–13762.

[7] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, “Cascade graph neural networks for rgb-d salient object detection,” in Proceedings of European Conference on Computer Vision, 2020, pp. 346–364.

[8] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, “A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection,” in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2020, pp. 9060–9069.

[9] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, “Cnn-based rgb-d saliency detection via cross-view transfer and multivision fusion,” IEEE Transactions on Cybernetics, vol. 48, no. 11, pp. 3171–3183, 2018.

[10] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, “Rgb-d salient object detection via 3d convolutional neural networks,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 2, 2021, pp. 1063–1071.

[11] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, “Metaformer is actually what you need for vision,” arXiv preprint arXiv:2111.11418, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proceedings of the International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” Proceedings of the International Conference on Learning Representations, 2021.

[14] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, “Tokens-to-token vit: Training vision transformers from scratch on imagenet,” arXiv preprint arXiv:2101.11986, 2021.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proceedings of European Conference on Computer Vision, 2020, pp. 213–229.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and Herv, “Training data-efficient image transformers & distillation through attention,” arXiv preprint arXiv:2012.12877, 2021.

[17] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, “Transform: Transformer-based object re-identification,” arXiv preprint arXiv:2102.04378, 2021.

[18] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, “Transformer tracking,” in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2021, pp. 8126–8135.

[19] H. Tan and M. Bansal, “Lxmer: Learning cross-modality encoder representations from transformers,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 5100–5111.

[20] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, “Tritransnet: Rgb-d salient object detection with a triplet transformer embedding
