DOI:10.19651/j.cnki.emt.2517812

基于深度网络特征交互的 RGB-T 显著目标检测*

魏明军[1,2] 杨 轩[1] 葛一珲[1] 刘亚志[1,2] 李 辉[1]
(1.华北理工大学人工智能学院 唐山 063210;2.河北省工业智能感知重点实验室 唐山 063210)

摘 要:针对现有方法中多模态间的互补信息利用不充分、特征交互易引入噪声的问题,提出了一种深度特征交互网络。首先,在编码阶段提出了深度特征多层交互模块,使用深度特征作为特征交互的线索,以充分利用可见光的纹理信息和热成像的位置信息。其次,设计了纹理位置特征交互模块,通过纹理信息与位置信息进行交互以实现同层级间的特征互补。然后,在解码阶段提出了膨胀卷积特征融合模块,通过膨胀卷积块提高模型感受野,使模型关注网络中的多尺度信息。最后,在公共 RGB-T 数据集 VT5000、VT1000、VT821 进行了广泛实验,实验表明,所提出网络的平均绝对误差分别达到 2.2%、1.5%、2.5%,与领域内先进的方法相比,取得了优异的性能。

关键词:显著目标检测;多模态;特征融合;RGB-T;特征交互;深度学习
中图分类号:TP391.41;TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.6040

RGB-T salient object detection based on deep network feature interaction

Wei Mingjun[1,2] Yang Xuan[1] Ge Yihui[1] Liu Yazhi[1,2] Li Hui[1]
(1. College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China;
2. Hebei Provincial Key Laboratory of Industrial Intelligent Perception, Tangshan 063210, China)

Abstract: A deep feature interaction network is proposed to address the problems of insufficient utilization of complementary information between multimodalities and the tendency of feature interaction to introduce noise in existing methods. First, a deep feature multilayer interaction module is proposed in the coding stage, which uses depth features as cues for feature interaction to fully utilize the texture information of visible light and the position information of thermal imaging. Second, a texture-position feature interaction module is designed to interact texture information with position information to achieve feature complementarity between the same layers. Then, the inflated convolutional feature fusion module is proposed in the decoding stage, which improves the model sensory field by inflating the convolutional block so that the model focuses on the multi-scale information in the network. Finally, extensive experiments are conducted on the public RGB-T datasets VT5000, VT1000 and VT821, which show that the average absolute errors of the proposed networks reach 2.2%, 1.5% and 2.5%, respectively, and achieve excellent performance compared with the state-of-the-art methods in the field.

Keywords: salient object detection; multimodal; feature fusion; RGB-T; feature interaction; deep learning

0 引 言

显著目标检测(salient object detection,SOD)旨在各种不同的场景中定位并分离图像和视频中最明显的目标。现已在各种计算机视觉研究中起到重要作用,例如图像分割、图像检索、目标跟踪、场景分类和动作检测等。近年来,显著目标检测算法成果众多,为了获取更精确的检测结果,显著目标检测的研究主流逐渐向基于深度学习的方法发展。例如,Qiao 等[1]提出了一种基于边缘分支和主体分支的双分支网络的 SOD 方法,该方法通过在边缘分支引入 4 个不同膨胀率的膨胀卷积层,并集成到一个模块中使得模型在不显著增加计算复杂度的情况下提高了边缘精度。在主体分支和边缘分支间通过交互编码器将两个分支的特征进行融合后使用双线性上采样来映射反馈给两个分支,实现更加精准的显著目标检测。Shen 等[2]为了提高显著目标检测的性能,提出了一种基于超像素单元的尺度缩减池化方法和相应的解池化方法,通过利用超像素信息计算对应单元表示的特征值。通过激活计算来搜索超像素单元最

收稿日期:2025-01-04
*基金项目:河北省教育厅重点课题(ZD2022102)、河北省工业智能感知重点实验室项目(SZX2021013)资助

具代表性的特征值,使模型更具鲁棒性,提高了检测结果的边界精度。Ullah 等[3]提出了一个复杂的上下文感知中间层引导网络,通过深度卷积大核与并行通道关注和压力动力注意力机制相结合,使模型能够在复杂的场景中有效地感知多尺度物体。作者通过使用深度卷积的大尺寸内核从邻接层到中心层范式中提取上下文和详细信息,丰富了模型捕获更多结构和上下文信息的能力,有效的增强了目标分割的精确度。朱磊等[4]为了增强上下文信息的关联,把 CNN 形式的自适应注意力和掩码注意力集成到网络中,通过交叉注意力和上下文特征增强模块,使网络聚焦于显著目标的整体区域,提升了模型的性能和泛化能力。朱书勤等[5]针对 RGB-D 多模态图像网络融合不充分、检测效率不高等问题,提出一种基于注意力机制的特征逐级融合网络结构。首先在使用两条 yolov3 骨干网络训练 RGB 和 Depth 图像,通过注意力模块增强两种特征,在网络中期逐层融合得到最终的特征权重。

然而,随着光照亮度降低、背景杂色混乱、显著目标边界信息模糊等极端问题的出现,仅依靠可见光(RGB)无法获取物体的位置信息和轮廓信息,导致模型的检测性能降低。得益于热像仪、RGB 相机在图像模态、感知精度和响应速度方面具有异质和互补的特性,热红外传感器可以将物体的热辐射信息补充出来。利用热红外传感器对物体进行成像,能够为可见光图像补充一定的轮廓信息和位置信息,因此 RGB-热成像显著目标检测(RGB-T SOD)成为近些年的研究热点课题。

目前,用于 RGB-T SOD 的方法有两种,一种是传统的基于图学习和排序的方法,但是这些手工提取的特征无法将像素之间的语义信息联系起来。另一种是基于深度学习的方法,其可以从预训练的骨干网络中提取单模态特征,通过各自包含的图像显著性线索进行信息互补,使用跨模态特征融合进行整合,最大限度的提升 RGB-T SOD 检测精度。Wu 等[6]提出了一种多模态融合网络。在编码器部分通过结合 RGB 和热成像特征,将两种模态的特征互补增强,实现跨模态融合。由浅层特征提供纹理信息,深层特征提供语义信息,在解码阶段采用自上而下的结构利用深层特征获得空间注意图,然后用空间注意图来引导低层注意图锐化预测边界,使模型关注显著区域。Lyu 等[7]提出一个跨模态注意力增强网络,通过使用交叉注意力单元来增强两个模态的特征,然后使用通道注意力来动态加权和融合两个模态的特征。Jin 等[8]认为人类本身识别显著目标主要以可见光为主体,特征融合时,热成像信息对 RGB 信息产生一定干扰,所以利用 RGB 为主要线索,热成像特征为次要线索,提出了一种新的跨模态非对称交互网络增强 RGB 和热成像的特征,在特征融合阶段在通道和空间两部分进行选择和融合,以寻求更合理的方式增强和融合多模态特征。葛荣泽等[9]针对可见光与红外图像融合算法中存在特征不平衡与特征融合不充分等问题,使用双流主干架构,提出了一种可见光-红外图像行人检测网络 MIFNet,并设计模态间信息融合模块,改变网络的结构减少特征不平衡造成的影响。孙铁强等[10]构建双流主干网络结构并行提取多模态信息和 RGB 信息,同时设计了一种即插即用的双模态特征交互修正融合模块提高信息交互能力。

然而这些深度学习方法实现的特征交互更侧重于同层级的特征,忽略了深层级特征所具有的语义信息,并且直接将两模态的特征进行拼接或融合引入了一定的噪声,使得跨模态交互过于粗糙,损害模型的判别能力。

考虑到上述问题,本文提出了一种使用深度特征作为线索,通过交叉注意力机制,引导同层级间特征交互的深度特征多层交互网络(deep features interaction net, DFINet),以 Swin Transformer 为基础框架,分别对 RGB 和热成像图像进行初步特征提取,在编码器中通过深度特征多层交互模块(deep features interaction module, DFI)将深层特征的语义信息作为上下文线索,在特征交互时提供上下文信息,增强 RGB 和热成像之间的互补能力,并使用纹理位置特征交互模块进一步增强模型对不同模态的互补能力。在解码器中使用膨胀卷积特征融合模块(dilatation convolution feature fusion, DCF)来提高模型感受野,并通过嵌入混合膨胀卷积减弱网格效应,最后逐层解码输出最终预测结果。

1 总体网络架构

如图 1 所示,本文的网络整体采用双流编码与解码结构。在编码器中,使用基于 Swin Transformer 的双编码网络模型,分别对可见光和热成像图像进行 4 个阶段的特征提取,前 3 个阶段使用深度特征多层交互模块(DFI),以深度特征 $\boldsymbol{F}_4^r$ 和 $\boldsymbol{F}_4^t$ 为线索,引导同层级间的特征进行特征交互。第 4 阶段使用纹理位置特征交互模块(texture position feature interaction module, TPM)实现多模态特征互补,以达到 RGB 和热成像特征互补增强的目的。在解码阶段,使用交互后的特征作为可见光解码器(R-Decoder)和热模态解码器(T-Decoder)的输入,分别输出热成像预测图 Out^{tir}、RGB 预测图 Out^{rgb},将预测图进行叠加,输入到膨胀卷积特征融合模块(DCF)中,增强模型对多尺度物体的关注度。模型包含 3 个预测结果,模型分别将 RGB、热成像解码器的输出,以及二者的预测图进行叠加,得到最终的显著性图。

2 相关工作

2.1 基于 Swin Transformer 的编码网络

受限于卷积核感受野尺寸,卷积层在处理图像数据时关注于局部区域的信息,无法直接获取图像的全局信息,这限制了模型在全局特征上的处理能力。Swin Transformer 是一种基于 Transformer 结构的深度学习模型,专门为图像识别任务而设计。模型的输入图像大小为 384×384,首

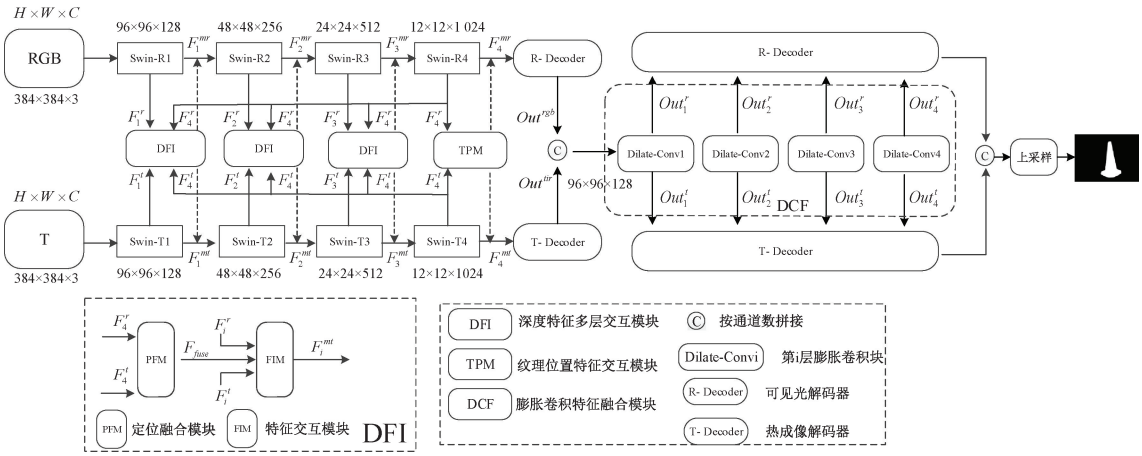

图 1　DFINet 模型总体架构

Fig. 1　DFINet network architecture

先将图像分割成一系列不重叠的小块，模型为分层结构，共分为 5 个阶段。每个阶段都是先进行下采样(其中第 1 阶段通过卷积层实现，其余 4 个阶段通过 Patch Merging 模块)，接下来通过 Swin Transformer block 利用窗口注意力和滑动窗口注意力对窗口内以及相邻窗口间分别进行自注意力计算，利用相邻窗口间的信息，对全局语义更好的建模。

本文使用基于 Swin Transformer 训练的 Swin-B 作为预训练模型，设计了双编码网络模型，利用不同阶段提取出的不同尺度的特征图进行交互和融合，因为底层特征的空间分辨率较大，且低级特征对于深度网络聚合性能贡献较小，为了减小模型计算量，只保留 Swin Transformer 的后 4 个阶段的输出。

2.2　深度特征多层交互模块

为了解决深层语义特征利用不充分、特征交互过于粗糙、易引入噪声的问题，受文献[11]启发，设计了 DFI 模块。

DFI 模块如图 1 所示，其中包含定位融合模块(positioning fusion module，PFM)和特征交互模块(future interaction module，FIM)，因为将深度特征直接融合易引入一定程度的噪声，RGB 特征虽具有丰富的纹理信息，但网络过深也会导致位置信息丢失。为了增强 RGB 特征的位置信息，减弱噪声，设计了定位融合模块，然后通过特征多层交互模块提高模型对于跨模态特征的学习能力，利用融合特征引导其他浅层特征进行特征交互，使其目标相关的对象和环境上下文信息相互补充。

PFM 模块如图 2 所示，首先深度特征 $\boldsymbol{F}_4{}^r$ 和 $\boldsymbol{F}_4{}^t$ 通过 1×1 卷积和归一化层进行拼接，公式如下：

$$\boldsymbol{F}_{add} = BN(\mathrm{Conv1}(\boldsymbol{F}_4^r)) + BN(\mathrm{Conv1}(\boldsymbol{F}_4^t)) \tag{1}$$

式中：Conv1 表示 1×1 卷积，BN 表示 Batch Normalization 操作，+表示相加操作。

然后通过大核注意力机制，因其将大核卷积从普通卷积、扩张卷积、点卷积 3 个方面进行了分解，使得降低了大核卷积计算量的同时可以捕获 RGB、热成像特征间的长距离关系。上述公式表示为：

$$\boldsymbol{F}_{lka} = \boldsymbol{F}_{add} \otimes \mathrm{Conv1}\begin{pmatrix}\mathrm{DW\_D\_Conv}\\(\mathrm{DW\_C\_Conv}(\boldsymbol{F}_{add}))\end{pmatrix} \tag{2}$$

式中：⊗表示逐元素相乘。DW_C_Conv 表示大核深度卷积，DW_D_Conv 表示大核膨胀卷积。

接下来通过跳跃连接使得模型更好的学习 RGB 和热成像特征之间的纹理信息、位置信息的差异，最后与包含更多位置信息的热成像特征 $\boldsymbol{F}_4{}^t$ 进行逐元素相乘，为模型补充位置信息和深度特征中的语义信息，实现特征间的定位融合。公式表示为：

$$\boldsymbol{F}_{fuse} = \boldsymbol{F}_4^t \otimes Sigmod(BN(\boldsymbol{F}_{lka} + \boldsymbol{F}_{add})) \tag{3}$$

式中：$Sigmod$ 表示激活函数。

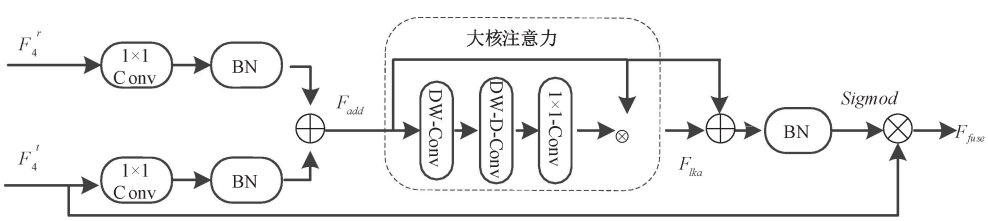

图 2　定位融合模块

Fig. 2　Positioning fusion module

FIM 基于交叉注意力实现,如图 3 所示。

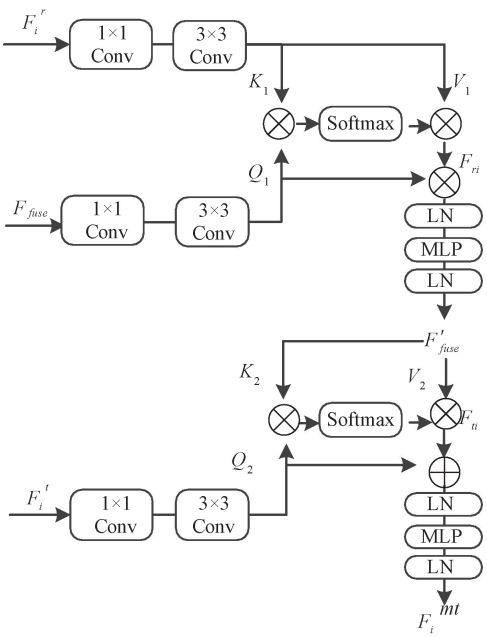

图 3 特征多层交互模块

Fig. 3 Feature multilayer interaction module

首先输入特征 $\boldsymbol{F}_i^r$、$\boldsymbol{F}_{fuse}$、$\boldsymbol{F}_i^t$ 都通过 1×1 和 3×3 卷积,改变特征的通道数,接下来使用 3 个不同的权重矩阵对特征图 $\boldsymbol{F}_{fuse}$,$\boldsymbol{F}_i^r$ 进行线性变换,变换后的特征图,分别表示为 $\boldsymbol{Q}_1$、$\boldsymbol{K}_1$、$\boldsymbol{V}_1$。

$$\begin{cases}\boldsymbol{Q}_1 = \boldsymbol{F}_{fuse}\boldsymbol{W}_q^1\\ \boldsymbol{K}_1 = \boldsymbol{F}_i^r\boldsymbol{W}_k^1\\ \boldsymbol{V}_1 = \boldsymbol{F}_i^r\boldsymbol{W}_v^1\end{cases} \tag{4}$$

式中:$\boldsymbol{W}_q^1$、$\boldsymbol{W}_k^1$、$\boldsymbol{W}_v^1$ 分别表示查询、键、值的权重矩阵。

然后 $\boldsymbol{Q}_1$ 和 $\boldsymbol{K}_1$ 的转置进行逐元素相乘,通过 $Softmax$ 层与 $\boldsymbol{V}_1$ 相乘,得到 $\boldsymbol{F}_{ri}$:

$$\boldsymbol{F}_{ri} = Softmax\left(\frac{\boldsymbol{Q}_1\boldsymbol{K}_1^{\mathrm{T}}}{\sqrt{C}}\right)\boldsymbol{V}_1 \tag{5}$$

之后,将 $\boldsymbol{F}_{ri}$ 与 $\boldsymbol{F}_{fuse}$ 相加,通过 LN 层和 MLP 层,使得特征中包含 RGB 中的目标上下文信息以及融合后的多模态目标的先验信息,这些信息对热成像目标特征起到增强作用,以上操作表示为:

$$\boldsymbol{F}'_{fuse} = LN\begin{pmatrix}LN(\boldsymbol{F}_{fuse}+\boldsymbol{F}_{ri})+\\ MLP(LN(\boldsymbol{F}_{fuse}+\boldsymbol{F}_{ri}))\end{pmatrix} \tag{6}$$

式中:LN 表示归一化层,MLP 表示全连接层。

同式(4),将 $\boldsymbol{F}_i{}^t$、$\boldsymbol{F}'_{fuse}$ 再次进行线性变换得到特征图 $\boldsymbol{Q}_2$、$\boldsymbol{K}_2$、$\boldsymbol{V}_2$。

$$\begin{cases}\boldsymbol{Q}_2 = \boldsymbol{F}'_{fuse}\boldsymbol{W}_q^2\\ \boldsymbol{K}_2 = \boldsymbol{F}_i^t\boldsymbol{W}_k^2\\ \boldsymbol{S}_2 = \boldsymbol{F}_i^t\boldsymbol{W}_v^2\end{cases} \tag{7}$$

最后,将 $\boldsymbol{F}_{ti}$ 与 $\boldsymbol{F}_i^t$ 相加,同样通过 LN 层和 MLP 层,以增强两个模态对应的目标相关区域的交互能力,得到交互后的特征 $\boldsymbol{F}_i^{mt}$。

$$\boldsymbol{F}_i^{mt} = LN(LN(\boldsymbol{F}_i^t+\boldsymbol{F}_{ti})+MLP(LN(\boldsymbol{F}_i^t+\boldsymbol{F}_{ti}))) \tag{8}$$

在相反的方向上,使用相同的方法,获得增强的 RGB 特征 $\boldsymbol{F}_i^{mr}$。

2.3 纹理位置特征交互模块

因为深度特征交互模块中的 PFM 模块可以提高 RGB 和热成像特征的交互能力,所以在此使用类似 PFM 模块的 TPM 模块,加强热成像特征的纹理细节,使多模态特征进行充分互补。

如图 4 所示,首先深度特征 $\boldsymbol{F}_4{}^r$ 和 $\boldsymbol{F}_4{}^t$ 通过卷积和归一化层进行拼接,表示为 $\boldsymbol{F}_{add}$,如式(1)所示,然后通过大核注意力机制,表示为 $\boldsymbol{F}_{lka}$,如式(2)所示。

接下来通过跳跃连接,最后与包含更多纹理信息的 RGB 特征 $\boldsymbol{F}_4{}^r$ 进行逐元素相乘,为热模态补充纹理信息,实现纹理信息与位置信息的交互。公式表示为:

$$\boldsymbol{F}_4^{mt} = \boldsymbol{F}_4^r \otimes Sigmod(BN(\boldsymbol{F}_{lka}+\boldsymbol{F}_{add})) \tag{9}$$

由于在 DFI 模块中已经为 RGB 特征补充了热成像特征的位置信息和深度语义信息,为了减小模型的计算量,故不再进行二次计算,将式(3)中的特征 $\boldsymbol{F}_{fuse}$ 作为特征交互后的结果 $\boldsymbol{F}_4{}^{mr}$。

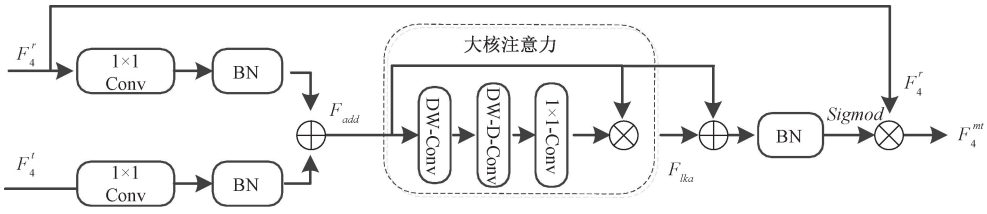

图 4 纹理位置特征交互模块

Fig. 4 Texture position feature interaction module

2.4 膨胀卷积特征融合模块

为了能够让模型拥有更大的感受野,关注到多尺度信息,提高对不同大小物体的检测精度,本文设计了膨胀卷积特征融合模块(DCF)。如图 5 所示,DCF 由 4 部分的膨胀卷积层(Dilate-Conv1~ Dilate-Conv4)组成,将膨胀卷积层级联,以进一步将两种模态的信息进行聚合,每层膨胀卷积分别输出到对应的 RGB 解码器(R-Decoder)和热模态解码器(T-Decoder)中逐步解码。

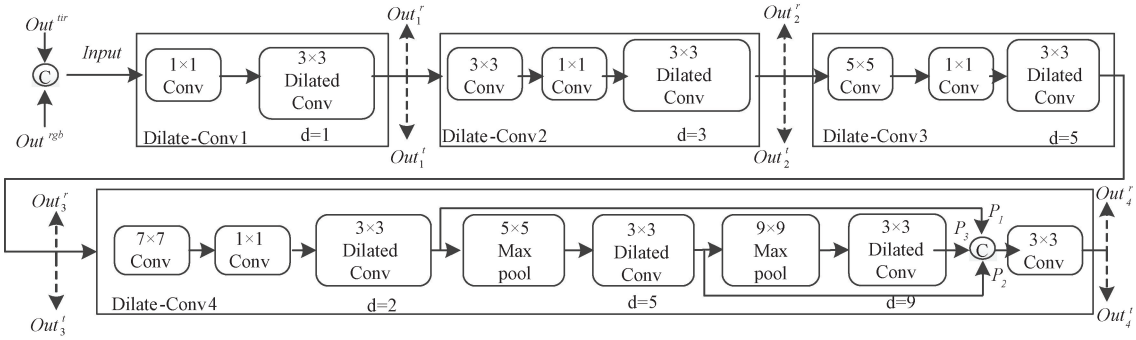

图 5　膨胀卷积特征融合模块

Fig. 5　Dilatation convolution feature fusion module

首先将两分支的预测图 Out_{tir}、Out_{rgb} 拼接作为 DCF 的输入，公式表示为：

$$\boldsymbol{In\ put} = \mathrm{Concat}(\boldsymbol{Out}^{rgb}, \boldsymbol{Out}^{tir}) \tag{10}$$

模型在学习多尺度特征时，低级特征对全局语义信息的贡献度较低，所以前 3 个膨胀卷积层中的膨胀率设置为 $r=1、3、5$，且膨胀卷积核大小为 3×3，在每个膨胀卷积块前设置 $r\times r$ 和 1×1 的普通卷积层，每一个膨胀卷积层后都使用 stride=2 的最大池化。上述公式表示为：

$$\boldsymbol{Out}_1 = \mathrm{Conv}_{r=1}(\mathrm{Conv1}(\boldsymbol{In\ put})) \tag{11}$$

$$\boldsymbol{Out}_1^t, \boldsymbol{Out}_1^r = \boldsymbol{Out}_1 \tag{12}$$

式中：Conv$i$ 表示 $i\times i$ 卷积，$\mathrm{Conv}_{r=i}$ 表示膨胀率为 i 的膨胀卷积，为简化公式，每层膨胀卷积的输出表示为 Out_i。

$$\boldsymbol{Out}_2 = P_{\max}(Conv_{r=3}(Conv1(Conv3(\boldsymbol{Out}_1)))) \tag{13}$$

$$\boldsymbol{Out}_2^t, \boldsymbol{Out}_2^r = \boldsymbol{Out}_2 \tag{14}$$

式中：$P_{\max}$ 同样表示最大池化。

$$\boldsymbol{Out}_3 = P_{\max}(\mathrm{Conv}_{r=5}(\mathrm{Conv1}(\mathrm{Conv5}(\boldsymbol{Out}_2)))) \tag{15}$$

$$\boldsymbol{Out}_3^t, \boldsymbol{Out}_3^r = \boldsymbol{Out}_3 \tag{16}$$

然而，基于膨胀卷积的架构会出现网格效应的问题。由于使用膨胀卷积时，因膨胀率不同而扩大了卷积核的接受域，需要在卷积核中相邻的权值间填充零权值。过多的零权值对于模型来说可能会丢失相邻信息，破坏信息的空间一致性，损害了模型提取多尺度深度特征的表征能力。所以在第 4 层膨胀卷积层中，使用混合膨胀卷积来捕获全局语义特征来减弱网格效应的问题，具体做法是设置了 3 个膨胀率分别为 $r_i=2、5、9$ 的膨胀卷积块，膨胀卷积完成后都进行 $r_{i+1}\times r_{i+1}$ 的最大池化层，$\boldsymbol{P}_1$、$\boldsymbol{P}_2$、$\boldsymbol{P}_3$ 分别表示通过最大池化层后的输出结果：

$$\boldsymbol{P}_1 = \mathrm{Conv7}(\mathrm{Conv1}(\mathrm{Conv}_{d=2}(\boldsymbol{Out}_3))) \tag{17}$$

$$\boldsymbol{P}_2 = \mathrm{Conv}_{d=5}(\mathrm{Maxpool}_{r=5}(\boldsymbol{P}_1)) \tag{18}$$

$$\boldsymbol{P}_3 = \mathrm{Conv}_{d=9}(\mathrm{Maxpool}_{r=9}(\boldsymbol{P}_2)) \tag{19}$$

式中：$\mathrm{Maxpool}_{r=}i$ 表示 stride=i 最大池化。这样即使发生网格效应涉及到了零权值，下一部分的卷积层也可以覆盖整个局部区域，不会丢失有用的相邻信息。

最后将三部分的深度特征图进行拼接，得到最终具有全局语义特征的特征图：

$$\boldsymbol{Out}_4^t, \boldsymbol{Out}_4^r = \mathrm{Conv3}(\mathrm{Concat}(\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3)) \tag{20}$$

将膨胀率设置为 2、5、9 是因为膨胀卷积块不能有大于 1 的公约数，最大池化层的设计也是为了避免膨胀卷积的网格效应，而将 r_3 中的膨胀率设置为 9 是要在最后一层膨胀卷积中让模型拥有更加广泛的感受野，以提高模块对于多尺度信息的表征能力。

2.5　损失函数

本文所设计的网络模型在解码器部分输出 RGB 预测图、热成像预测图、及二者拼接后的预测图三部分，所以损失函数包括热成像预测输出、RGB 预测输出、拼接后特征图预测输出，因此总损失函数 L 可以表示为：

$$L = L_{RGB} + L_{TIR} + L_{FUSE} \tag{21}$$

式中：L_{RGB}、L_{TIR}、L_{FUSE} 分别表示可见光的损失、热成像的损失、拼接融合后的损失。

融合后的特征具有可见光和热成像的全局结构，在这里使用 IOU 损失[12]衡量融合结构的相似性：

$$L_{IOU}(\boldsymbol{F}^r, \boldsymbol{F}^t) = \frac{1}{|C|}\sum_i^T \frac{\boldsymbol{F}_i^r \cdot \boldsymbol{F}_i^t}{\boldsymbol{F}_i^r + \boldsymbol{F}_i^t - \boldsymbol{F}_i^r \cdot \boldsymbol{F}_i^t} \tag{22}$$

式中：$\boldsymbol{F}^r = \{\boldsymbol{F}_i^r \mid i=1,\cdots,T\}$ 为网络中可见光特征图，$\boldsymbol{F}^t = \{\boldsymbol{F}_i^t \mid i=1,\cdots,T\}$ 表示网络热成像的特征图，T 是图像中的总像素数。

使用 BCE 损失[13]平衡像素之间的平滑梯度，BCE 损失函数定义如下：

$$L_{BCE}(\boldsymbol{S}, \boldsymbol{Y}) = -\sum_{i=1}^{T}\begin{pmatrix}\boldsymbol{Y}_i \log(\boldsymbol{S}_i) + (1-\boldsymbol{Y}_i) \\ \log(1-\boldsymbol{S}_i)\end{pmatrix} \tag{23}$$

式中：$\boldsymbol{S} = \{\boldsymbol{S}_i \mid i=1,\cdots,T\}$ 为网络预测的显著图，$\boldsymbol{Y} = \{\boldsymbol{Y}_i \mid i=1,\cdots,T\}$ 为真值图，T 是图像中的总像素数。

使用 SSIM 损失[14]促进像素值之间的相关性使得模型聚焦于显著目标的边界，因其公式较为复杂，具体公式如文献[14]所示。

L_{FUSE} 公式可以表示为三部分损失之和：

$$L_{FUSE} = L_{IOU} + L_{BCE} + L_{SSIM} \tag{24}$$

因为 L_{RGB}、L_{TIR} 不存在融合结构，直接取 BCE 损失和 SSIM 损失之和：

$$L_{RGB} = L_{BCE} + L_{SSIM} \tag{25}$$

$$L_{TIR} = L_{BCE} + L_{SSIM} \tag{26}$$

3 实 验

3.1 数据集

本文在 3 个 RGB-T SOD 数据集上对所提出的方法进行了评估。其中 VT821 包含 821 个手动注册的图像对,部分图像为了增加数据集的挑战性还添加了噪声。VT1000 包含 1 000 个 RGB-T 图像对,这些图像是由高度对齐的 RGB 和热像仪捕获的。VT5000 包含 5 000 对高分集、低偏差的 RGB-T 图像,其场景更为复杂,其中 3 个数据集分辨率都为 640×480,训练集和测试集的比例为 1:1。

在 3 个通用数据集中,所采集的图像是由不同光照亮度下的各种场景、物体所组成,并且包含小目标、动态场景、热交叉、添加噪声等特殊场景。本文在 3 个数据集中进行了定量实验和定性实验,在定性实验中,尤其对具有挑战的特殊场景与领域内先进方法进行了比对,以全面验证模型的泛化能力和先进性。

为了公平比较,使用相同的 VT5000 中 2500 对图像的训练数据集进行训练,对 VT1000、VT821、VT5000 中其余图像用于网络的测试评估。

3.2 实施细节

本文网络基于 Pytorch 在单个显卡型号为 NVIDIA GeForce RTX 3090(24 GB 显存)GPU 和型号为 24 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz 的 CPU 的计算机上运行实现。

实验过程中,两个骨干网的参数由 ImageNet22k 上预训练的 Swin-B 初始化,两分支间不共享权重,epoch 设置为 200,学习率设置为 0.001,动量设置为 0.9。在模型的训练阶段,批量大小设置为 8,将图像输入大小设置为 384×384。

3.3 评估指标

采用领域内广泛使用的指标来评估模型和 SOTA RGB-T SOD 模型的性能。它们分别是加权 F 度量(F_β)、结构测度 S 度量(S_α)、平均增强匹配标准 E 度量(E_ϵ)和平均绝对误差(MAE)来对所提出的网络进行评估。

结构测度 S 度量可以评估显著性图与真值之间的区域感知和对象感知结构相似性,公式如下:

$$S_\alpha = \alpha \times S_0 + (1-\alpha) \times S_r \tag{27}$$

式中:S_r 为区域感知的结构相似性,S_o 为对象感知的结构相似性,α 设置为 0.5。

加权 F 度量是精确率(Precision)和召回率(Recall)的加权调和平均值,公式表示如下:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN} \tag{28}$$

$$F_\beta = \frac{(1+\eta^2) \cdot Precision \cdot Recall}{\eta^2 \cdot Precision + Recall} \tag{29}$$

式中:TP、FP 和 FN 分别表示真阳性、假阳性和假阴性,并且 η^2 被设置为 0.3。

E-measure[15]用于全局统计信息和局部像素匹配信息,公式如下:

$$E_\epsilon = \frac{(2 \cdot Precision \cdot Recall)}{Precision + Recall} \tag{30}$$

平均绝对误差(mean absolute error, MAE)测量显著性图和真值图之间的每像素绝对差的平均值,公式表示为:

$$MAE = \frac{1}{T}\sum_{i=1}^{T} |\boldsymbol{S}_i - \boldsymbol{Y}_i| \tag{31}$$

在实验中 E-measure 和 F-measure 采用自适应值。在模型评估时 F_β、S_α、E_ϵ 越大越好,MAE 越小越好。

3.4 实验结果与分析

1)定量对比

为了验证本模型 DFINet 的优越性,本文与近些年的 RGB-T SOD SOTA 算法进行了比较。如表 1 所示,其中包括:SwinNet、MITF、CAE、CMDB、FAWT、MMHL、RGBTS、CMIM、CAFC,最优值以加粗格式标出。

表 1 定量对比实验结果

Table 1 Quantitative comparison experimental results

数据集	评估指标	SwinNet[16]	MTFI[17]	CAE	CMDB[18]	FAWT[19]	MMHL[20]	RGBT[21]	CMIM[22]	CAFC	本文
VT821	S_α	0.904	0.905	0.884	0.878	0.883	0.892	0.857	0.893	0.891	**0.915**
	F_β	0.847	0.853	0.820	0.841	0.802	0.830	0.835	0.846	0.856	**0.856**
	E_ϵ	0.926	0.927	0.916	0.923	0.907	0.923	0.904	0.920	0.927	**0.936**
	MAE	0.030	0.027	0.036	0.031	0.033	0.029	0.036	0.031	0.028	**0.025**
VT1000	S_α	0.938	0.938	0.923	0.923	0.933	0.929	0.906	0.931	0.935	**0.943**
	F_β	0.896	0.906	0.881	0.901	0.863	0.893	0.899	0.903	0.912	**0.919**
	E_ϵ	0.947	0.949	0.949	0.961	0.937	0.941	0.939	0.943	0.951	**0.968**
	MAE	0.018	0.016	0.023	0.021	0.021	0.021	0.027	0.018	0.017	**0.015**
VT5000	S_α	0.912	0.910	0.880	0.879	0.884	0.886	0.843	0.888	0.899	**0.921**
	F_β	0.865	0.870	0.823	0.846	0.807	0.823	0.817	0.851	0.869	**0.878**
	E_ϵ	0.942	0.943	0.919	0.932	0.910	0.926	0.903	0.923	0.939	**0.954**
	MAE	0.026	0.025	0.038	0.032	0.036	0.033	0.042	0.034	0.027	**0.022**

本文在 3 个通用数据集中进行了定量实验对比，由实验可知，DFINet 网络在多个评价指标中优于以上 9 个先进方法，证明了 DFINet 的优越性且所选用的损失函数可以帮助网络更好的检测 RGB-T 显著目标。

2)定性对比

由于一些算法的显著性分割图未公开，因此在图 6 中展示了与定量实验不完全一致的先进模型的定性对比结果，其中包括不同场景下具有挑战的性能比较。

例如交叉边界下的多显著对象(第 1 行、第 5 行)、动态模糊(第 2 行)、多显著对象(第 3 行)、高亮度物体(第 4 行)、中心偏移的小目标显著对象(第 6 行)、高噪声下的显著对象(第 7 行)、灰度处理下的显著对象(第 8 行)、热交叉(第 9 行)、中心偏移的大物体显著对象(第 10 行)等。

结果表明，本文所提出的模型在应对各种挑战场景时表现出更好的检测效果，这得益于模型充分利用了深度特征中的多模态上下文信息，将其与 RGB 图像中的纹理信息和热成像图像中的位置信息进行充分互补，特征融合模块增大了模型对多尺度物体的感受野，提高了模型对多尺度信息的挖掘能力，实现了更精确的 RGB-T SOD。

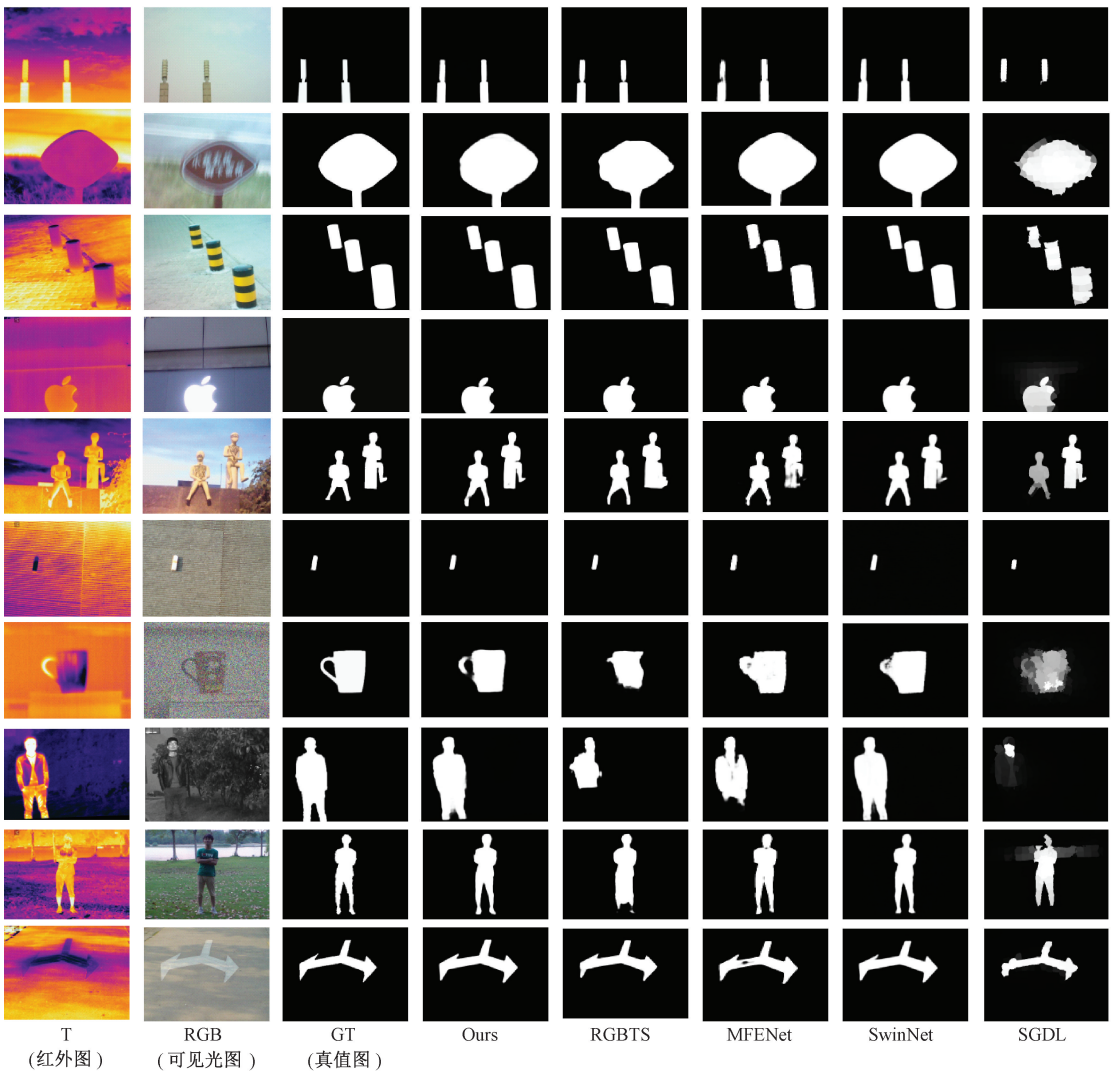

图 6　定性对比实验结果

Fig. 6　Qualitative comparison experimental results

3)消融实验

为了研究深度特征交互网络中不同模块对网络性能的影响，本文首先在保留其他模块的同时去除了 DFI 模块，如表 2 所示，相比完整的深度特征指导的特征多层交互网络，当去掉 DFI 时(第 1 行)，由于网络无法将 RGB 特征的纹理信息和热成像特征的位置信息互补，造成检测性能降低。

当使用不添加深度特征为线索的特征交互模块时(第 2 行)，因为缺失了深度特征所包含的语义信息、PFM 模块所补充的位置信息，RGB 特征和热成像特征无法充分互

补,导致检测精度降低。当去掉 DCF 模块,使用普通 3×3 卷积代替时(第 3 行),模型提取多尺度信息能力下降,导致模型对尺度大小、形状不一的物体检测精度降低。

表 2 DFI 网络消融实验结果

Table 2 Results of DIFNet ablation experiment

模块	VT5000			
	S_α	F_β	E_ϵ	MAE
W/O DFI	0.914	0.862	0.947	0.024
W/O PFM	0.918	0.874	0.951	0.022
W/O DCF	0.915	0.868	0.949	0.024
本文	**0.921**	**0.878**	**0.954**	**0.022**

4)复杂度分析

Params 是模型中所包含的参数数量,FLOPs (floating point operations) 指模型的计算量,可以用来衡量算法或模型的复杂度,FPS (frames per second) 是模型每秒可以处理的帧数,常用于实时任务。不同算法的 Params、FLOPs 和 FPS 如表 3 所示。

本文对比了两种不同方法的 RGB-T 网络,一种是同样基于 Swin Transformer 的 SwinNet,相比而言 SwinNet 的参数量更高,FLOPS 也同样高于本模型,这是因为 SwinNet 在解码器部分使用了大量的卷积、上采样的聚合模块,导致计算量上升。另一种是基于 CNN 的 CGFNet[23]。由于 CGFNet 使用卷积实现,导致模型复杂度上升,本文所提出的方法在复杂度部分相较 CGFNet 表现更好。

为了评估模型的实用性,本文在模型的推理速度方面进行了实验,由于 SwinNet 和 CGFNet 并未提供相关 FPS 数据,为保证数据可靠性,本文在相同的实验环境中对两个模型进行了实验,由表 3 可知,本文所使用的方法高于 CGFNet,略低于 SwinNet。

表 3 复杂度分析结果

Table 3 Complexity analysis results

算法	Params/M	FLOPs/G	FPS
SwinNet	198.70	124.30	29
CGFNet	66.38	139.97	16
本文	175.30	97.70	26

4 结 论

本文提出一种深度特征交互网络(DFINet),为了更好的利用热成像与 RGB 之间的互补信息,设计了深度特征多层交互模块,将不同模态的深度特征融合,作为多模态特征交互的线索提供语义信息,提高模型对多模态特征间的纹理信息、位置信息学习能力。为使模型更好的关注到多尺度物体,设计了膨胀卷积的特征融合模块,在级联的膨胀卷积层中嵌入了混合膨胀卷积,以减弱网格效应。网络还使用纹理位置特征交互模块将纹理信息、位置信息进行同层间特征交互,增强多模态互补能力。

大量的实验表明,DFINet 相较主流的先进模型,具有更好的 RGB-T SOD 检测结果。但本文也存在复杂度高的问题,这是因为所提出的交互模块中需要计算双向交叉注意力导致,在未来,将会考虑模型的轻量化创新,使用计算量相对较低的方法实现特征交互,并将该模型运用到实际应用中,在模型精度与推理速度之间寻求平衡。

参考文献

[1] QIAO M, ZHOU G, LIU Q L, et al. Salient object detection: An accurate and efficient method for complex shape objects[J]. IEEE Access, 2021, 9: 169220-169230.

[2] SHEN C H, CHEN Y J, HSIAO H F. A scale-reductive pooling with majority-take-all for salient object detection [C]. 2022 IEEE International Symposium on Circuits and Systems(ISCAS). IEEE, 2022: 3309-3313.

[3] ULLAH I, HUSSAIN S, SHAHEED K, et al. CMGNet: Context-aware middle-layer guidance network for salient object detection[J]. Journal of King Saud University-Computer and Information Sciences, 2024, 36(1): 101838.

[4] 朱磊,袁金垚,王文武,等.自适应卷积注意力与掩码结构协同的显著目标检测[J].电子与信息学报,2024,47(3):1-10.
ZHU L, YUAN J Y, WANG W W, et al. Significant object detection with adaptive convolutional attention and mask structure coordination[J]. Journal of Electronics and Information Technology, 2024, 47(3): 1-10.

[5] 朱书勤.基于注意力融合网络的 RGB-D 目标检测算法[J].电子测量技术,2021,44(9):110-115.
ZHU SH Q. RGB-D object detection algorithm based on attention fusion network [J]. Electronic Measurement Technology, 2021, 44(9): 110-115.

[6] WU J Y, ZHOU W J, QIAN X H, et al. MFENet: Multitype fusion and enhancement network for detecting salient objects in RGB-T images[J]. Digital Signal Processing, 2023, 133: 103827.

[7] LYU C, WAN B, ZHOU X, et al. CAE-Net: Cross-modal attention enhancement network for RGB-T salient object detection[J]. Electronics, 2023, 12(4): 953.

[8] JIN D Z, SHAO F, XIE ZH X, et al. CAFCNet: Cross-modality asymmetric feature complement network for RGB-T salient object detection[J]. Expert

Systems with Applications, 2024, 247: 123222.

[9] 葛荣泽,武一.基于阶段特征融合的图像融合行人检测[J].电子测量技术,2024,47(24):103-109.
GE R Z, WU Y. Image fusion pedestrian detection based on stage feature fusion [J]. Electronic Measurement Technology, 2024, 47(24): 103-109.

[10] 孙铁强,魏光辉,宋超,等.基于 YOLO 的多模态钢轨表面缺陷检测方法[J].电子测量技术,2024,47(21):72-81.
SUN T Q, WEI G H, SONG CH, et al. Multi modal rail surface defect detection method based on YOLO [J]. Electronic Measurement Technology, 2024, 47(21): 72-81.

[11] HUI T R, XUN Z ZH, PENG F G, et al. Bridging search region interaction with template for RGB-T tracking [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 13630-13639.

[12] MATTYUS G, LUO W, URTASUN R. Deep road mapper: extracting road topology from aerial images[C]. IEEE International Conference on Computer Vision (ICCV). IEEE,2017: 3438-3446.

[13] BOER P T D, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method[J]. Annals of operations research, 2005,134:19-67.

[14] WANG ZH, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]. 37th IEEE Asilomar Conference on Signals, Systems and Computers, 2003:1398－1402 .

[15] FAN D P, CHENG G, CAO Y, et al. Enhanced-alignment measure for binary foreground map evaluation [J]. ArXiv preprint arXiv: 1805.10421, 2018.

[16] LIU ZH Y, TAN Y CH, HE Q, et al. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(7): 4486-4497.

[17] CHEN G, SHAO F, CHAI X L, et al. Modality-induced transfer-fusion network for RGB-D and RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(4):1787-1801.

[18] XIE ZH X, SHAO F, CHEN G, et al. Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8):4149-4163.

[19] ZHANG P, XU M N, ZHANG Z Y, et al. Feature aggregation with transformer for RGB-T salient object detection[J]. Neurocomputing, 2023, 546: 126329.

[20] REN G X, JOSHI J, CHO Y. Multi-modal hybrid learning and sequential training for RGB-T saliency detection[J]. ArXiv preprint arXiv:2309.07297,2023.

[21] LIU ZH Y, HUANG X SH, ZHANG G H, et al. Scribble-supervised RGB-T salient object detection[C]. 2023 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2023: 2369-2374.

[22] LYU CH T, WAN B, ZHOU X F, et al. Lightweight cross-modal information mutual reinforcement network for RGB-T salient object detection [J]. Entropy, 2024, 26(2): 130.

[23] WANG J, SONG K CH, BAO Y Q, et al. CGFNet: Cross-guided fusion network for RGB-T salient object detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32 (5): 2949-2961.

作者简介

魏明军(通信作者),教授,硕士,主要研究方向为计算机视觉、入侵检测、机器学习和数据挖掘。

E-mail: weimj@ncst.edu.cn

杨轩,硕士研究生,主要研究方向为计算机视觉、深度学习。

E-mail: 1056320360@qq.com

葛一珲,硕士研究生,主要研究方向为计算机视觉。

E-mail: 1195831022@qq.com

刘亚志,教授,博士,主要研究方向为计算机视觉、移动计算。

E-mail: liuyazhi@ncst.edu.cn

李辉,本科,副教授,主要研究方向为数据挖掘、计算机视觉。

E-mail: 13091076301@163.com