DOI:10.19652/j.cnki.femt.2305809

# 基于跨模态特征融合的 RGB-D 显著性目标检测

李可新 何 丽 刘哲凝 钟润豪
(新疆大学智能制造现代产业学院(机械工程学院) 乌鲁木齐 830017)

**摘 要**:RGB-D 显著性目标检测因其有效性和易于捕捉深度线索而受到越来越多的关注。现有的工作通常侧重于通过各种融合策略学习共享表示,少有方法明确考虑如何维持 RGB 和深度的模态特征。提出了一种跨模态特征融合网络,该网络维持 RGB-D 显著目标检测的 RGB 和深度的模态,通过探索共享信息以及 RGB 和深度模态的特性来提高显著检测性能。具体来说,采用 RGB 模态、深度模态网络和一个共享学习网络来生成 RGB 和深度模态显著性预测图以及共享显著性预测图。提出了一种跨模态特征融合模块,用于融合共享学习网络中的跨模态特征,然后将这些特征传播到下一层以整合跨层次信息。此外,提出了一种多模态特征聚合模块,将每个单独解码器的模态特定特征整合到共享解码器中,这可以提供丰富的互补多模态信息来提高显著性检测性能。最后,使用跳转连接来组合编码器和解码器层之间的分层特征。通过在 4 个基准数据集上与 7 种先进方法进行的实验表明,方法优于其他最先进的方法。

**关键词**:RGB-D 显著性目标检测;跨模态融合网络;跨模态特征融合;多模态聚合
**中图分类号**: TN2 **文献标识码**:A **国家标准学科分类代码**: 520.604

# RGB-D salient object detection based on cross-modal feature fusion

Li Kexin He Li Liu Zhening Zhong Runhao
(College of Intelligent Manufacturing Modern Industry (College of Mechanical Engineering), Xinjiang University, Urumqi 830017, China)

**Abstract**: RGB-D saliency object detection has received increasing attention due to its effectiveness and ease of capturing depth cues. Existing work usually focuses on learning shared representations through various fusion strategies, and few approaches explicitly consider how to maintain the modal features of RGB and depth. In this paper, we propose a cross-modal fusion network that maintains the modalities of RGB and depth for RGB-D salient object detection, and improves the salient detection performance by exploring the shared information as well as the properties of RGB and depth modalities. Specifically, an RGB modal, a deep modal network, and a shared learning network are used to generate RGB and deep modal saliency prediction maps as well as shared saliency prediction maps. A cross-modal feature integrate module is proposed to fuse cross-modal features in the shared learning network, which are then propagated to the next layer for integrating cross level information. Besides, we propose a multi-modal feature aggregation module to integrate the modality specific features from each individual decoder into the shared decoder, which can provide rich complementary multi-modal information to boost the saliency detection performance. Further, a skip connection is used to combine hierarchical features between the encoder and decoder layers. Experiments with ten state-of-the-art methods on four benchmark datasets show that the method in this paper outperforms other state-of-the-art methods.

**Keywords**: RGB-D saliency object detection; cross modal fusion network; cross modal feature integrate module; multi-modal feature aggregation

**收稿日期**:2023-12-03

## 0 引 言

显著性目标检测(SOD)的目的是通过模仿人类视觉注意的特征来识别图像中最显著的区域或物体。已被广泛应用于计算机视觉相关任务,如图像理解、视频/语义分割、动作识别和行人再识别[1-4]。尽管取得了重大进展,但在复杂场景中准确定位显著物体仍然具有挑战性,例如实例杂乱的背景或低对比度的光照条件。最近,随着智能设备中深度传感器的广泛应用,引入了深度图来提供几何和空间信息,以提高显著检测性能。因此,融合RGB和深度图像在显著检测社区中引起了越来越多的关注。

对于RGB-D显著性目标检测,融合RGB和深度图像至关重要。现有的RGB-D显著性目标检测的融合策略,其大致可以分为3类,早期融合、后期融合和多尺度融合[5]。早期融合策略使用简单的连接方式来整合RGB和深度图。Song等[6]直接将RGB和深度图像整合在一起,形成一个四通道输入。然而,这种类型的融合没有考虑两种模态之间的分布差距,这可能会导致不准确的特征融合。基于后期融合策略的各种模型使用两个并行网络流来为RGB和深度数据生成独立的显著性图,然后将两个图像融合一起,以获得最终的预测图[7-8]。然而这种类型的融合很难捕捉到两种模态之间的复杂相关性。多尺度融合策略利用两个独立的网络分别学习两种模态的中间特征,然后将融合的特征输入后续网络或解码器。该策略可以有效的使网络学习特征之间的相关性。Zhao等[9]在基于CNNs的架构中引入对比度先验来增强深度信息,然后使用流体金字塔集成模块将增强的深度与RGB特征进行集成。孙福明等[10]通过采用一种双向交互融合的方式,将融合特征作为RGB特征的补充,以降低低质量深度图带来的负面影响。上述方法主要关注融合共享表示来进行学习,然后使用解码器生成最终的显著图。此外,没有监督解码器来指导基于深度的特征学习,这可能会妨碍获得最佳的深度特征。从多模态学习的角度来看,一些工作[11-12]表明,探索共享信息和RGB或深度模态特征可以提高模型性能。然而,少有RGB-D显著性检测模型明确利用RGB或深度模态的特征。

为此,本文提出了一种基于跨模态特征融合的RGB-D显著性目标检测网络。该网络不仅探索了共享信息,还利用了RGB或深度模态的特征来提高显著性检测性能。在该模型中,使用两个编码器子网络来提取两种模态的多尺度特征,并提出了跨模态特征集成模块(cross-modal feature integrate module, CFIM)来融合跨模态特征。然后,使用U-Net[13]结构构建了一个多模态解码器,其中编码器和解码器层之间的跳转连接用于组合分层特征。该模型可以在每个独立的解码器中学习强大的模态特征。除此之外,本文还构建了一个共享解码器,使用跳转连接将来自先前CFIM的分层特征组合在一起。为了充分利用RGB或深度模态的特征,利用多模态特征聚合(multi-modal feature aggregation, MFA)将其集成到共享解码器中。最后,制定了一个统一的端到端可训练框架来实现RGB-D显著性目标检测。

本文提出一个基于跨模态特征融合的RGB-D显著性目标检测网络。既能探索共享特征,又能维持RGB和深度模态的特异性。本文提出了一个CFIM来融合RGB或深度模态特征并学习两种模态的共享特征。每个CFIM的输出传输到下一层以捕捉多尺度信息。本文利用一个简单而有效的MFA模块来聚合学习到的RGB或深度模态特征。充分利用在解码器中学习到的特征来提高显著性目标检测性能。根据4种常用的评价指标,本文提出的方法在4个公共权威RGB-D数据集上均取得优秀的成绩,从而证明其先进的性能。

## 1 本文模型

### 1.1 模型整体框架

基于跨模态特征融合的RGB-D显著性目标检测网络的框架如图1所示。首先,使用两个编码器结构,将RGB-D和深度图像输入到编码器中,分别学习RGB-D和深度图像的特定模态特征,并且提出了CFIM来学习它们的共享特征表示。然后,采用RGB、深度和共享解码器子网络分别生成显著性预测图。此外,编码器网络的原始特征通过跳转连接集成到解码器中。为了充分利用RGB或深度模态解码器学习的特征,本文提出了一个MFA模块将这些特征集成到共享解码器中,以提高显著目标检测性能。

如图1所示,使用Res2Net-50[14]建立子网络,并在ImageNet数据集上进行了预训练。其中,解码器子网络中有5个多层特征,即RGB模态输出特征 $F^R=[f_m^R, m=1,2,\cdots,5]$ 和深度模态输出特征 $F^D=[f_m^D, m=1,2,\cdots,5]$。本文将特定模态编码器子网络的输入分辨率表示为 $W\times H$。因此,对第1层的特征分辨率设置为 $W/8\times H/8$,一般分辨率为 $W/2m\times H/2m(m>1)$。此外,设第 $m$ 层的特征通道数为 $C_m(m=1,2,\cdots)$,则 $C=[64,256,512,1\,024,2\,048]$。

当得到高阶特征 $F_5^R$ 和 $F_5^D$,会送入解码器子网络中,产生单独的显著图。然后,本文利用U-Net结构来构建多模态解码器,其中编码器和解码器之间的跳转连接被用来结合分层特征。在解码器阶段各层输出的特征 $F^R$ 和 $F^D$,与解码器阶段对应层次的特征图进行拼接,并采用采用了适当的卷积核大小和步长,并引入了批归一化和激活函数来加速训练和提高模型的表达能力。除此之外,连接特征(在解码器子网络的首层只有 $F_5^R$ 和 $F_5^D$)送入感受场模块(reception field block,RFB),以捕捉全局的上下文信息。RGB和深度模态的学习网络能够有效地学习其特定的属性,这些特征被整合到共享解码器子网络中以提高显著性目标检测的性能。

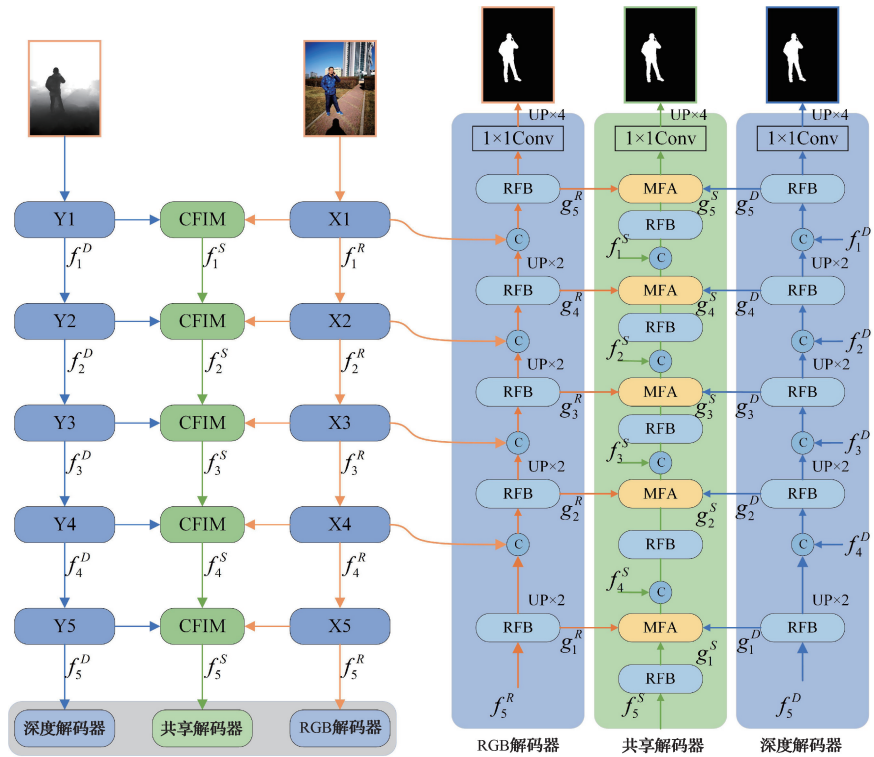


图 1 跨模态融合网络模型

Fig. 1 Cross-modal fusion network model

如图 1 所示，在共享网络中，本文模型融合了 RGB 和深度模态的交叉模态特征来学习共享特征，并将其输入共享解码器以生成最终的显著图。然后，本文还在编码器和解码器层之间采用跳跃连接结合分层特征。此外，为了充分利用多模态解码器学习过的特征，将其整合到共享解码器中，以提高显著性目标检测性能。

## 1.2 跨模态特征融合模块

本文提出用于高效融合跨模态特征的 CFIM。CFIM 模块包括两部分，跨模态特征增强和自适应特征融合，如图 2 所示。首先将第 $i$ 层的 RGB 特征 $F_i^R = \boldsymbol{R}^{W_i \times H_i \times C_i}$ 和深度特征 $F_i^D = \boldsymbol{R}^{W_i \times H_i \times C_i}$（其中 $\boldsymbol{R}$ 表示特征集合，$W$、$H$、$C$ 表示第 $i$ 层特征的高度、宽度和通道数）通过 $1 \times 1$ 的卷积层使得通道数减少至 $C_i/2$，以获得加速效果。使用交叉增强策略，通过学习两种模态的增强特征来利用它们之间的相关性。RGB 特征和深度特征被送至具有 Sigmoid 激活函数的 $3 \times 3$ 卷积层中，然后可以获得归一化的特征图 $w_i^R$ 和 $w_i^D$：

$$w_i^R = \sigma(Conv_3(F_i^D)) \in [0,1] \tag{1}$$

$$w_i^D = \sigma(Conv_3(F_i^R)) \in [0,1] \tag{2}$$

式中：$Conv_3()$ 表示 $3 \times 3$ 卷积操作，$\sigma()$ 表示 Sigmloid 激活函数。为了充分利用两种模态之间的相关性，归一化的特征图可以被视为特征级注意力图，以自适应地增强特征表示。此外，为了保留每个模态的原始信息，残差连接适

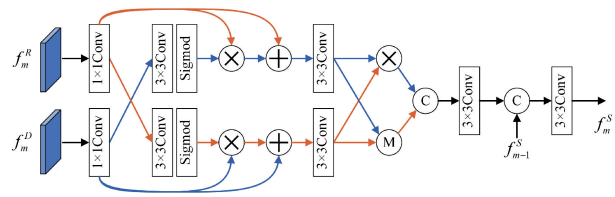


图 2 跨模态特征融合模块示意图

Fig. 2 Diagram of the proposed cross-modal feature integrate module

用于将增强的特征与其原始特征相结合。因此，两种模态的交叉增强特征表示 $F_i^{R'}$ 和 $F_i^{D'}$：

$$F_i^{R'} = F_i^R + F_i^R \otimes w_i^D \tag{3}$$

$$F_i^{D'} = F_i^D + F_i^D \otimes w_i^R \tag{4}$$

式中：$\otimes$ 表示元素相乘，$i \in \{1,2,3,4,5\}$。

使用自适应融合策略，将获得的交叉增强特征表示（$F_i^{R'}$ 和 $F_i^{D'}$）有效地融合它们。应用元素乘法和最大化，然后将结果连接在一起。将两个交叉增强特征表示被馈送到卷积层中，获得平滑表示，然后进行逐个元素相乘和最大化。因此，可以获得：

$$p_{mul} = Bconv_3(F_i^{R'}) \otimes Bconv_3(F_i^{D'}) \tag{5}$$

$$p_{\max} = Max(Bconv_3(F_i^{R'}), Bconv_3(F_i^{D'})) \tag{6}$$

式中：$Bconv_3(\cdot)$ 表示结合 $3 \times 3$ 卷积层和 $ReLu$ 函数的顺序操作；$Max()$ 表示最大化操作；$p_{mul}$ 和 $p_{\max}$ 表示逐个元

素相乘和最大化的结果。然后，将逐个元素相乘和最大化的结果连接为 $p_{cat}=[p_{mul},p_{\max}]$，通过 $Bconv_3(\cdot)$ 操作获得 $p_{cat}^1=Bconv_3(p_{cat})$，结合第 $i-1$ 层 CFIM 的输出结果 $F_{i-1}{}^S$ 送至 $Bconv_3(\cdot)$ 操作，最后得到共享特征 $F_i^S$：

$$F_i^S=Bconv_3([Bconv_3(p_{cat}),F_{i-1}^S]) \quad i\in\{2,3,4,5\} \tag{7}$$

$$F_i^S=Bconv_3(Bconv_3(p_{cat})) \quad i=1 \tag{8}$$

式中：$Bconv_3(\cdot)$ 表示结合 3×3 卷积层和 *ReLu* 函数的顺序操作。

本文提出的 CFIM 可以通过交叉增强特征学习有效地利用两种模态之间的相关性，并通过自适应加权不同的特征表示来融合它们。此外，融合的特征 $F^S$ 被传播到下一层，以捕获和融合多层信息。

### 1.3 多模态聚合模块

为了充分利用在特定模态解码器中学到的特征，本文提出了一种简单但有效的 MFA 模块，将其集成到共享解码器中。在共享解码器的第 $m$ 层，可以得到共享特征表示 $g_m^S$，以及特定模态特征 $g_m^R$ 和 $g_m^D$。如图 3 所示，特征 $g_m^R$ 和 $g_m^D$ 与当前层共享特征 $g_m^S$ 相乘，得到 $g_m^{RS}=g_m^S\otimes g_m^R$ 和 $g_m^{DS}=g_m^S\otimes g_m^D$。这两个特征进一步连接（$[g_m^{RS},g_m^{DS}]$），然后送入 $Bconv(\cdot)$ 运算，得到特征 $g_m^{Sc}$。最后，通过加法运算将卷积特征 $g_m^{Sc}$ 与原始特征 $g_m^S$，得到 MFA 的输出结果。

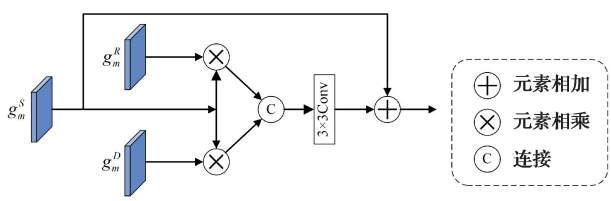


图 3 多模态聚合模块示意图

Fig. 3 Diagram of the proposed multi-modal feature aggregation module

MFA 的输出结果是将解码器阶段学习到 RGB 和深度特征进行融合的结果，将其集成到共享解码器中，实现增强共享表征，并提供丰富而互补的跨模态信息。除此之外，RGB 和深度模态解码器可获得监督信号，以指导特征学习，从而保留多模态的属性，这有利于将最终预测结果整合到共享解码器中。

### 1.4 损失函数

最后，本文制定了一个统一的终端可训练框架。整体功能由两部分组成，即 $L_{sp}$ 和 $L_{sp}$，分别用于多模态解码器和共享解码器。为方便起见，$S_R$ 和 $S_D$ 分别表示使用 RGB 和深度图像时的预测图，$S_{sh}$ 表示使用其共享表示的预测图，G 表示真实值。因此，总体损失函数可表述如下：

$$L_{total}=L_{sh}(S_{sh},G)+L_{sp}(S_R,G)+L_{sp}(S_D,G) \tag{9}$$

其中，利用像素位置感知损失函数用于 $L_{sp}$ 和 $L_{sh}$，可以对难识别和易识别的像素给予不同的关注，从而提高显著性目标检测的性能。

## 2 实验结果与分析

为了验证所提模型的有效性，在 4 个具有挑战性 RGB-D 数据集上对其进行了评估，包括 NLPR[15] 由微软 Kinect 拍摄的 1 000 组 RGB 和深度图像组成；NJUD[16] 包含 2 003 个具有不同目标和复杂场景的立体图像；SIP[17] 包含 929 幅突出人物的高分辨率图像；STERE[18] 包含从互联网收集的 1 000 组双目图像，是该领域第一个立体图像数据集。为了让对比更加公平，本文采用了和以往工作相同[6-12]的方式，训练集从 NJU2K 中选取了 1 485 个样本，从 NLPR 中选取了 700 个样本，总共选取了 2 195 个样本进行训练。其余的 NJUD 和 NLPR 样本，以及整个 SIP 和 STERE 样本将用于测试。为了定量评估本文网络和其他 RGB-D 显著性目标检测方法的结果，采用了 5 种广泛使用的评估指标来分析不同方法的性能，召回率曲线（Precision-Recall），F-measure（$F_\beta$）[19]，平均绝对误差（MAE）[20]，S-measure（$S_\alpha$）[21]，E-measure（$E_\xi$）[22]。

本文模型基于 PyTorch 实现，并在配备 16 GB 内存的 Nvidia Quard RTX 5000 GPU 上进行训练。使用的骨干网络（Res2Net-50）已在 ImageNet 上预先训练过。由于 RGB 和深度图像的通道不同，深度编码器的输入通道被修改为 1。本文采用 Adam 算法来优化所提出的模型。初始学习率设为 $1\times10^{-4}$，权重衰减为每 60 轮衰减为原先的 0.1 倍。RGB 和深度图像的输入分辨率被调整为 352×352。使用随机翻转、旋转和边界剪切等多种策略对训练图像进行增强。批次大小设置为 20，模型训练轮次为 200。在测试阶段，RGB 和深度图像被调整为 352×352 大小，然后输入模型以获得预测图。然后，再将预测图重新调整为原始大小，以实现最终评估。最后，共享解码器的输出就是该模型的最终预测图。

本文在 4 个数据集上，使用 4 种评价指标，将本文提出的模型与目前具有代表性的 10 种深度学习方法进行性能对比，其中包括 DF[23]、CTMF[24]、PCF[25]、D3Net[17]、SSF[26]、ICNet[27]、S2MA[28]、BBS-Net[29]、DCF[30] 和 UC-Net[31]。为了与其他方法公平对比，本文使用作者提供的显著性图分别在定性上和定量上进行对比。

### 2.1 定量对比

如表 1 所示，使用 4 个评估指标在 4 个公共数据集上进行实验对比的结果，其中“↑”表示该列的评价指标值越大越好，“↓”表示该列的评价指标越小越好，字体黑色加粗代表该值为当列最优值，字体灰色加粗代表该值为当列次优值。本文模型明显优 7 种先进方法。此外，在 NLPR 和 SIP 数据集上，本文方法在 4 项评估指标上都优于所有同类最先进的方法，并获得了最佳性能。与其他 RGB-D 显著性目标检测方法相比，本文模型在 NJU2K 和 STERE 数据集上获得了较好的性能，并且与 UCNet 相媲美。总

之，本文提出的模型在给定场景中定位显著性目标方面取得了较好的成绩。此外，图 4 和 5 所示为 *P*-*R* 曲线和 F-measure 曲线。由 7 种 RGB-D 显著性目标检测方法的结果可以看出，本文的模型获得了更好的结果，生成的曲线大部分领先于其他模型的优越性较为明显。总之，定量结果充分证明了本文所提出模型的有效性和优越性。

**表 1　与其他方法在 4 个数据集上的定量比较结果**

**Table 1　Quantitative comparison results with other methods on four datasets**

| 数据集 | 指标 | DF | PCF | CTMF | SSF | ICNet | S2MA | D3Net | BBS | DCF | UCNet | 本文 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NLPR | $F_\beta$ ↑ | 0.759 2 | 0.794 9 | 0.723 5 | 0.861 8 | 0.870 0 | 0.901 7 | 0.896 8 | 0.917 9 | 0.896 1 | 0.903 8 | 0.918 5 |
| | $S_\alpha$ ↑ | 0.805 9 | 0.873 6 | 0.859 9 | 0.885 1 | 0.922 6 | 0.915 6 | 0.911 8 | 0.930 5 | 0.912 4 | 0.917 4 | 0.918 2 |
| | $E_\xi$ ↑ | 0.883 7 | 0.916 3 | 0.869 1 | 0.933 2 | 0.943 9 | 0.953 2 | 0.952 9 | 0.960 9 | 0.952 8 | 0.957 0 | 0.961 6 |
| | $M$ ↓ | 0.078 9 | 0.043 7 | 0.056 1 | 0.035 0 | 0.028 1 | 0.030 2 | 0.029 8 | 0.023 4 | 0.026 3 | 0.025 5 | 0.020 8 |
| MJUD | $F_\beta$ ↑ | 0.783 4 | 0.843 9 | 0.787 6 | 0.900 4 | 0.868 0 | 0.837 3 | 0.832 7 | 0.919 7 | 0.906 8 | 0.910 3 | 0.927 7 |
| | $S_\alpha$ ↑ | 0.768 2 | 0.876 8 | 0.849 3 | 0.931 4 | 0.894 0 | 0.882 8 | 0.873 7 | 0.938 1 | 0.938 0 | 0.952 4 | 0.943 2 |
| | $E_\xi$ ↑ | 0.839 3 | 0.896 3 | 0.863 7 | 0.938 4 | 0.904 5 | 0.905 8 | 0.898 5 | 0.949 3 | 0.943 4 | 0.949 3 | 0.957 0 |
| | $M$ ↓ | 0.136 0 | 0.059 2 | 0.084 7 | 0.042 1 | 0.051 9 | 0.066 0 | 0.066 8 | 0.035 1 | 0.039 0 | 0.035 0 | 0.038 6 |
| SIP | $F_\beta$ ↑ | 0.673 3 | 0.824 6 | 0.683 5 | 0.785 6 | 0.835 9 | 0.877 1 | 0.861 0 | 0.883 5 | 0.872 6 | 0.892 5 | 0.903 6 |
| | $S_\alpha$ ↑ | 0.652 9 | 0.842 4 | 0.715 8 | 0.798 7 | 0.853 8 | 0.872 1 | 0.860 3 | 0.878 9 | 0.862 7 | 0.875 7 | 0.893 6 |
| | $E_\xi$ ↑ | 0.794 3 | 0.898 8 | 0.823 9 | 0.870 0 | 0.898 8 | 0.918 2 | 0.908 5 | 0.921 5 | 0.914 9 | 0.925 3 | 0.932 8 |
| | $M$ ↓ | 0.185 4 | 0.070 6 | 0.139 4 | 0.051 3 | 0.069 5 | 0.057 5 | 0.063 3 | 0.054 8 | 0.058 4 | 0.049 3 | 0.043 0 |
| STERE | $F_\beta$ ↑ | 0.742 2 | 0.826 4 | 0.770 8 | 0.839 5 | 0.864 7 | 0.882 2 | 0.891 1 | 0.903 2 | 0.897 8 | 0.908 4 | 0.906 3 |
| | $S_\alpha$ ↑ | 0.757 4 | 0.874 6 | 0.848 0 | 0.836 9 | 0.902 5 | 0.890 4 | 0.898 5 | 0.908 3 | 0.900 6 | 0.900 4 | 0.909 8 |
| | $E_\xi$ ↑ | 0.838 2 | 0.896 7 | 0.864 3 | 0.871 9 | 0.915 3 | 0.912 6 | 0.919 4 | 0.928 8 | 0.937 4 | 0.937 8 | 0.942 9 |
| | $M$ ↓ | 0.140 9 | 0.063 5 | 0.086 3 | 0.064 7 | 0.044 6 | 0.051 4 | 0.046 2 | 0.041 1 | 0.037 7 | 0.039 2 | 0.038 9 |

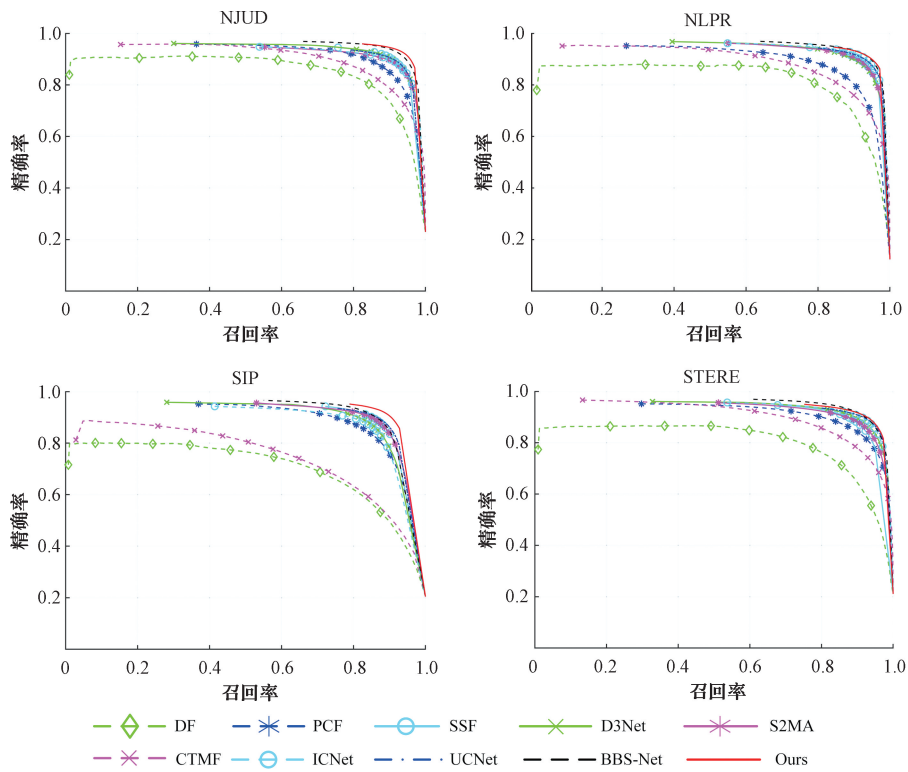


图 4　不同方法在 4 个数据集上的 *P*-*R* 曲线

Fig. 4　*P*-*R* curves of different methods on four datasets

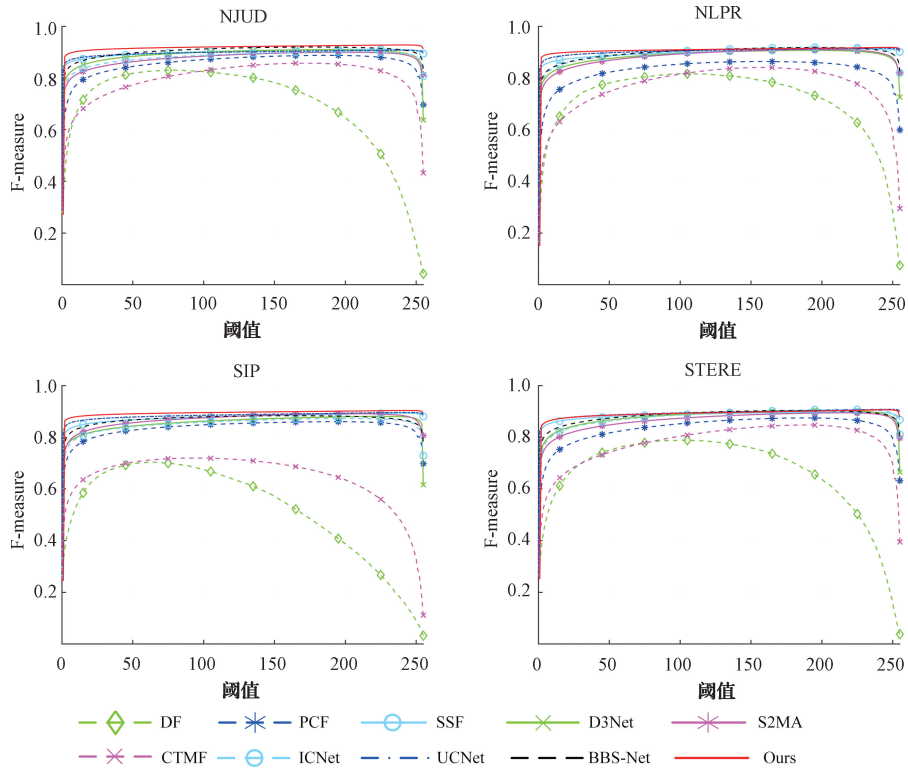


图 5　不同方法在 4 个数据集上的 F-measure 曲线

Fig. 5　F-measure curves of different methods on four datasets

## 2.2　定性对比

图 6 所示为本文模型与 7 种最先进方法进行比较的几个代表性结果。本文方法、DCF 和 D3Net 可以准确检测到显著性目标，而 BBS-Net、S2MA、SSF 和 UCNet 则预测了一些非物体区域。图 6B 和 C 两个复杂背景场景的例子。从对比结果可以看出，本文方法和 S2MA 能产生

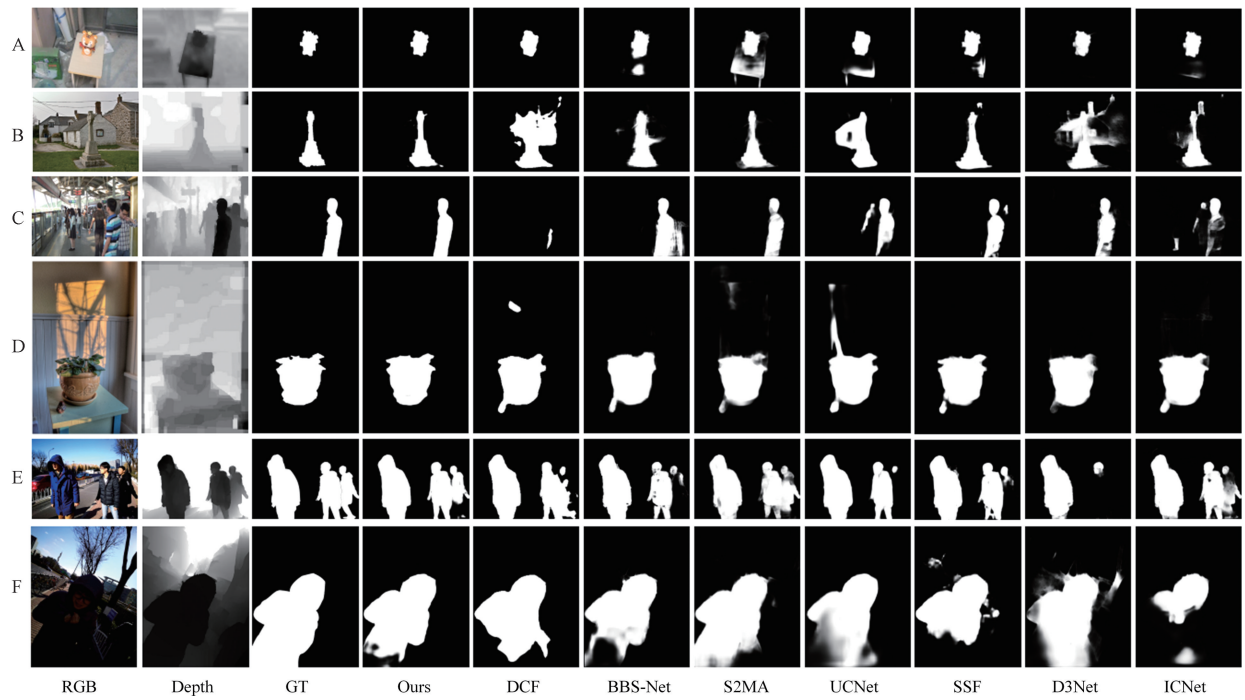


图 6　本文方法与 7 种最先进方法可视化对比

Fig. 6　Visual comparisons of our method and seven state-of-the-art methods

可靠的结果，而其他 RGB-D 显著性目标检测模型则无法定位物体或将背景混淆为突出物体。图 6D 是一个有多个显著目标的示例，可以看出，准确定位所有显著目标。与其他方法相比，本文方法能定位所有显著目标，并更准确地分割它们，生成更清晰的边缘。图 6F 一个弱光条件下的例子。有些方法无法检测到突出物体的整个范围。本文模型可以通过抑制背景干扰物来提高显著性检测性能，从而产生令人满意的结果。

### 2.3 消融实验

为了验证模型中不同组成部分的有效性，本文进行了消融研究。

1)跨模态特征融合模块的有效性

由于所提出的 CFIM 是用于融合跨模型特征并学习它们的共享特征，因此采用直接连接策略来代替 CFIM。本文将两个特征 $f_m^R$ 和 $f_m^R$（图 2)直接连接起来，然后输入到 3×3 卷积层，从而在每一层获得融合表示。表 2 中，将这一评估结果记为“A1”。从比较结果可以看出，本文模型在使用 CFIM 时比使用样本特征串联策略时表现更好。这也说明了 CFIM 在提高显著性检测性能方面的贡献。此外，CFIM 包括两个部分，即跨模型特征增强和自适应特征融合。因此，为了评估每个部分的贡献，将只进行了跨模型特征增强或自适应特征融合的 CFIM 分别记为“A2”和“A3”。将这两个独立的部分与完整版的 CFIM 进行比较，可以看到所提出的 CFIM 的有效性。为了验证传播策略的有效性，在 CFIM 中删除了这一传播，记为“A4”。“A4”和 CFIM 的对比结果表明，这种传播策略提高了显著性目标的检测性能。

**表 2 消融实验的定量评估**

**Table 2 Quantitative evaluation for ablation studies**

| | NJUD | | STERE | | NLPR | | SIP | |
|---|---|---|---|---|---|---|---|---|
| | $S_\alpha\uparrow$ | $M\downarrow$ | $S_\alpha\uparrow$ | $M\downarrow$ | $S_\alpha\uparrow$ | $M\downarrow$ | $S_\alpha\uparrow$ | $M\downarrow$ |
| 本文 | 0.926 | 0.028 | 0.907 | 0.037 | 0.927 | 0.021 | 0.894 | 0.043 |
| A1 | 0.916 | 0.034 | 0.898 | 0.042 | 0.926 | 0.022 | 0.892 | 0.044 |
| A2 | 0.921 | 0.031 | 0.895 | 0.042 | 0.925 | 0.022 | 0.896 | 0.042 |
| A3 | 0.919 | 0.032 | 0.895 | 0.043 | 0.929 | 0.020 | 0.887 | 0.048 |
| A4 | 0.924 | 0.029 | 0.903 | 0.038 | 0.927 | 0.023 | 0.888 | 0.046 |
| B1 | 0.918 | 0.034 | 0.901 | 0.041 | 0.922 | 0.024 | 0.885 | 0.048 |
| B2 | 0.924 | 0.029 | 0.900 | 0.041 | 0.926 | 0.022 | 0.893 | 0.044 |
| B3 | 0.921 | 0.031 | 0.903 | 0.039 | 0.925 | 0.022 | 0.891 | 0.045 |

2)多模态聚合模块的有效性

在本文提出的框架中，MFA 被提议充分利用在特定模式解码器中学习到的特征，然后将这些特征整合到共享解码器中，以提供更多的多模式互补信息。为了验证其有效性，我们删除了这一模块，记为“B1”。此外，还考虑将另外两种特征融合策略与 MFA 进行比较，一种是跨模态特征增强融合，另一种是简单的串联策略，分别称为“B2”和“B3”。如表 2 所示，将“B1”与完整模型进行比较，结果表明了所学特征整合到共享解码器中的有效性。将“B2”和“B3”与本文完整模型进行比较，可以看到 MFA 模块的效果优于其他两种融合策略。

## 3 结 论

本文提出一个基于跨模态特征融合的 RGB-D 显著性目标检测网络。与大多数主要关注学习共享特征的现有研究不同，本文模型不仅探索了共享的跨模态信息，还对 RGB 和深度模态特征进行了补偿，从而提高了显著性目标检测性能。然后，利用 CFIM 来融合 RGB 和深度模态特征并学习两种模态的共享特征，实现跨模态和跨尺度传播信息。最后 MFA 模块可以为共享解码器提供特定属性，以增强多模态信息的互补性。在 4 个具有挑战性的基准数据集上进行的定量和定性评估实验表明，本文模型优于现有的 RGB-D 显著性目标检测方法。

### 参 考 文 献

[1] 张晓宁，王雨青，陈小林. 基于多路径递归增强的显著性目标检测方法[J]. 国外电子测量技术，2021，40(5)：1-7.
ZHANG X N，WANG Y Q，CHEN X L. Multi-path recurrent enhanced salient object detection method[J]. Foreign Electronic Measurement Technology，2021，40(5)：1-7.

[2] 钱晓亮，张鹤庆，张焕龙，等. 基于视觉显著性的太阳能电池片表面缺陷检测[J]. 仪器仪表学报，2017，38(7)：1570-1578.
QIAN X L，ZHANG H Q，ZHANG H L，et al. Solar cell surface defect detection based on visual saliency[J]. Chinese Journal of Scientific Instrument，2017，38(7)：1570-1578.

[3] RAPANTZIKOS K，AVRITHIS Y，KOLLIAS S. Dense saliency-based spatiotemporal feature points for action recognition[C]. IEEE Conference. Computer Vision and Pattern Recognition. IEEE，2009：1454-1461.

[4] 项家伟，王伟. 基于显著性目标检测网络的面部属性编辑方法[J]. 国外电子测量技术，2022，41(5)：1-8.
XIANG J W，WANG W. Trans-to-editing-face attributes editing via salient object detection[J]. Foreign Electronic Measurement Technology，2022，41(5)：1-8.

[5] ZHOU T，FAN D P，CHENG M M，et al. RGB-D salient object detection：A survey[C]. Proceedings of Computational Visual Media，2021：37-69.

[6] SONG H，LIU Z，DU H，et al. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning[J].

IEEE Transaction on Image Processing, 2017, 26(9): 4204.

[7] LIU Z, SHI S, DUAN Q, et al. Salient object detection for RGB-D image by single stream recurrent convolution neural network[J]. Neurocomputing, 2019, 363: 46-57.

[8] REN J, GONG X J, YU L, et al. Exploiting global priors for RGB-D saliency detection[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, 2015: 25-32.

[9] ZHAO J X, CAO Y, FAN D P, et al. Contrast prior and fluid pyramid integration for RGBD salient object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 3922-3931.

[10] 孙福明,胡锡航,武景宇,等.跨模态交互融合与全局感知的RGB-D显著性目标检测[J].软件学报,2024,35(4):1899-1913.
SUN F M, HU X H, WU J Y, et al. RGB-D salient object detection based on cross-modal interactive fusion and global awareness[J]. Journal of Software, 2024,35(4):1899-1913.

[11] 刘译善,孙涵.基于特征增强的RGB-D显著性目标检测[J].计算机技术与发展,2023,33(11):28-34.
LIU Y SH, SUN H. Feature enhancement based RGB-D salient object detection[J]. Computer Technology and Development,2023,33(11):28-34.

[12] LU Y,WU Y,LIU B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 13376-13386.

[13] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]. Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015:234-241.

[14] QU L, HE S, ZHANG J, et al. RGBD salient object detection via deep fusion[C]. Proceedings of IEEE Transactions on Image Processing. IEEE Computer Society, 2017,26(5): 2274-2285.

[15] PENG H, LI B, XIONG W, et al. RGBD, Salient object detection: A benchmark and algorithms[C]. European Conference on Computer Vision. Cham: Springer, 2014: 92-109.

[16] JU R, GE L, GENG W, et al. Depth saliency based on anisotropic center-surround difference[C]. 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014: 1115-1119.

[17] FAN D P, LIN Z, ZHANG Z, et al. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(5): 2075-2089.

[18] PIAO Y, JI W, LI J, et al. Depth-induced multi-scale recurrent attention network for saliency detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7254-7263.

[19] ZHU J Y, WU J, XU Y, et al. Unsupervised object class discovery via saliency-guided multiple class learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(4): 862-875.

[20] PERAZZI F, KRÄHENBÜH P, PRITCH Y, et al. Saliency filters: Contrast based filtering for salient region detection[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 733-740.

[21] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: A new way to evaluate foreground maps[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 4548-4557.

[22] FAN D P, GONG C, CAO Y, et al. Enhanced-alignment measure for binary foreground map evaluation[J]. ArXiv Preprint, arXiv: 1805.10421, 2018.

[23] QU L, HE S, ZHANG J, et al. RGBD salient object detection via deep fusion[J]. Proceedings of IEEE Transactions on Image Processing. IEEE Computer Society, 2017,26(5): 2274-2285.

[24] HAN J, CHEN H, LIU N, et al. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion[J]. IEEE Transactions on Cybernetics, 2018, 48(8): 2171-2183.

[25] CHEN H, LI Y. Progressively complementarity-aware fusion network for RGB-D salient object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3051-3060.

[26] ZHANG M, REN W, PIAO Y, et al. Select, supplement and focus for RGB-D saliency detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 3472-3481.

[27] LI G, LIU Z, LING H. ICNet: Information conversion network for RGB-D based salient object detection[J]. IEEE Transactions on Image Processing, 2020, 29(12): 4873-4884.

[28] ZHAO X, ZHANG L, PANG Y, et al. A single

stream network for robust and real-time RGB-D salient object detection[C]. European Conference on Computer Vision. Springer, 2020:92-109.

[29] LIU N, ZHANG N, HAN J. Learning selective self-mutual attention for RGB-D saliency detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2020:13753-13762.

[30] FAN D P, ZHAI Y, BORJI A, et al. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network[C]. European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 275-292.

[31] JI W, LI J, YU S, et al. Calibrated RGB-D salient object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9471-9481.

## 作 者 简 介

李可新,硕士研究生,主要研究方向为计算机视觉、图像处理、深度学习。
E-mail:likexin_1124@163.com