E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)05-1252-13

论文引用格式:Ye X Y, Zhu L, Wang W W and Fu Y. 2024. RGB_D salient object detection algorithm based on complementary information interaction. Journal of Image and Graphics, 29(05):1252-1264(叶欣悦, 朱磊, 王文武, 付云. 2024. 互补特征交互融合的RGB_D实时显著目标检测. 中国图象图形学报, 29(05):1252-1264)[DOI:10.11834/jig.230583]

互补特征交互融合的RGB_D实时显著目标检测

叶欣悦,朱磊*,王文武,付云

武汉科技大学信息科学与工程学院,武汉　430081

摘　要:**目的**　通过融合颜色、深度和空间信息,利用RGB_D这两种模态数据的显著目标检测方案通常能比单一模态数据取得更加准确的预测结果。深度学习进一步推动RGB_D显著目标检测领域的发展。然而,现有RGB_D显著目标检测深度网络模型容易忽略模态的特异性,通常仅通过简单的元素相加、相乘或特征串联来融合多模态特征,如何实现RGB图像和深度图像之间的信息交互则缺乏合理性解释。为了探求两种模态数据中的互补信息重要性及更有效的交互方式,在分析了传统卷积网络中修正线性单元(rectified linear unit,ReLU)选通特性的基础上,设计了一种新的RGB和深度特征互补信息交互机制,并首次应用于RGB_D显著目标检测中。**方法**　首先,根据该机制提出了互补信息交互模块将模态各自的"冗余"特征用于辅助对方。然后,将其阶段式插入两个轻量级主干网络分别用于提取RGB和深度特征并实施两者的交互。该模块核心功能基于修改的ReLU,具有结构简单的特点。在网络的顶层还设计了跨模态特征融合模块用于提取融合后特征的全局语义信息。该特征被馈送至主干网络每个尺度,并通过邻域尺度特征增强模块与多个尺度特征进行聚合。最后,采用了深度恢复监督、边缘监督和深度监督3种监督策略以有效监督提出模型的优化过程。**结果**　在4个广泛使用的公开数据集NJU2K(Nanjing University 2K)、NLPR(national laboratory of pattern recognition)、STERE(stereo dataset)和SIP(salient person)上的定量和定性的实验结果表明,以Max F-measure、MAE(mean absolute error)以及Max E-measure共3种主流测度评估,本文提出的显著目标检测模型相比较其他方法取得了更优秀的性能和显著的推理速度优势(373.8帧/s)。**结论**　本文论证了在RGB_D显著目标检测中两种模态数据具有信息互补特点,提出的模型具有较好的性能和高效率推理能力,有较好的实际应用价值。

关键词:显著目标检测(SOD);RGB_D;深度卷积网络;互补信息交互;跨模态特征融合

RGB_D salient object detection algorithm based on complementary information interaction

Ye Xinyue, Zhu Lei*, Wang Wenwu, Fu Yun

School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan　430081, China

Abstract: **Objective** By fusing color, depth, and spatial information, using RGB_D data in salient object detection typically achieves more accurate predictions compared with using a single modality. Additionally, the rise of deep learning technology has further propelled the development of RGB_D salient object detection. However, existing RGB_D deep network models for salient object detection often overlook the specificity of different modalities. They typically rely on simple

收稿日期:2023-08-23;**修回日期**:2024-01-08;**预印本日期**:2024-01-15

*通信作者:朱磊　zhulei@wust.edu.cn

基金项目:国家自然科学基金项目(61873196,61502358)

Supported by:National Natural Science Foundation of China(61873196,61502358)

fusion methods, such as element-wise addition, multiplication, or feature concatenation, to combine multimodal features. However, the existing models of significant object detection in RGB_D deep networks often ignore the specificity of different modes. They often rely on simple fusion methods, such as element addition, multiplication, or feature joining, to combine multimodal features. These simple fusion techniques lack a reasonable explanation for the interaction between RGB and depth images. These methods do not effectively take advantage of the complementary information between RGB and depth modes nor do they take advantage of the potential correlations between them. Therefore, more efficient methods must be proposed to facilitate the information interaction between RGB images and depth images so as to obtain more accurate significant object detection results. To solve this problem, the researchers simulated the relationship between RGB and depth by analyzing traditional neural networks and linear correction units (ReLU) (e. g. , structures, such as constructed recurrent neural networks or convolutional neural networks). Finally, a new interactive mechanism of complementary information between RGB and depth features is designed and applied to RGB_D salient target detection for the first time. This method analyzes the correlations between RGB and depth features and uses these correlations to guide the fusion and interaction process. To explore the importance of complementary information in both modalities and more effective ways of interaction, we propose a new RGB and depth feature complementary information interaction mechanism based on analyzing the selectivity of ReLU in traditional convolutional networks. This mechanism is applied for the first time in RGB_D salient object detection. **Method** First, on the basis of this mechanism, a complementary information interaction module is proposed to use the "redundancy" characteristics of each mode to assist each other. Then, it is inserted into two lightweight backbone networks in phases to extract RGB and depth features and implement the interaction between them. The core function of the module is based on the modified ReLU, which has a simple structure. At the top layer of the network, a cross-modal feature fusion module is designed to extract the global semantic information of the fused features. These features are passed to each scale of the backbone network and aggregated with multiscale features via a neighborhood scale feature enhancement module. In this manner, not only local and scale sensing features can be captured but also global semantic information can be obtained, thus improving the accuracy and robustness of salient target detection. At the same time, three monitoring strategies are adopted to supervise the optimization of the model effectively. First, the accuracy of depth information is constrained by depth recovery supervision to ensure the reliability of depth features. Second, edge supervision is used to guide the model to capture the boundary information of important targets and improve the positioning accuracy. Finally, deep supervision is used to improve the performance of the model further by monitoring the consistency between the fused features and the real significance graph. **Result** By conducting quantitative and qualitative experiments on widely used public datasets (Nanjing University 2K(NJU2K), national laboratory of pattern recognition(NLPR), stereo dataset(STERE), and salient person(SIP)), the salient object detection model in this study shows remarkable advantages on three main evaluation measures: Max F-measure, mean absolute error(MAE), and Max E-measure. The model performed relatively well, especially on the SIP dataset, where it achieved the best results. In addition, the processing speed of the model remarkably improved to 373. 8 frame/s, while the parameter decreased to 10. 8 M. Compared with the other six methods, the proposed complementary information aggregation module remarkably improved in the effect of salient target detection. By using the complementary information of RGB and depth features and through the design of cross-modal feature fusion module, the model can better capture the global semantic information of important targets and improve the accuracy and robustness of detection. **Conclusion** The proposed salient object detection model in this study is based on the design of complementary information interaction module, lightweight backbone network, and cross-modal feature fusion module. The method maximizes the complementary information of RGB and depth features and achieves remarkable performance improvement through optimized network structure and monitoring strategy. Compared with other methods, this model shows better results in terms of accuracy, robustness, and computational efficiency. In RGB_D data, this work is of crucial to deepening the understanding of the importance of multimodal data fusion and promoting the research and application in the field of salient target detection.

Key words: salient object detection(SOD); RGB_D; deep convolutional network; complementary information interaction; cross-modal feature fusion

0 引 言

随着深度学习的不断发展，显著目标检测（salient object detection，SOD）成为计算机视觉领域的研究热点之一（孙涵 等，2023）。其主要研究目的是寻找图像中最具有显著性的区域。具体包括从图像或者视频中找到具有显著性或者重要性的目标区域，即人眼在观看图像或视频时最先注意的目标区域，并将这些区域与背景分离出来。在SOD的探索实验中，研究人员发现，引入深度信息可以弥补RGB图像缺失的深度信息，从而有效检测出显著目标（丛润民 等，2023）。这种新方法称为RGB_D SOD，它利用了RGB图像和深度图像，融合了两者的优点。早期的模型依赖于手工提取特征，然后将其融合在一起，但由于获取深度图像的难度较大，所以这一领域的研究一直在缓慢进行。随着Microsoft Kinect等深度传感器的使用，深度图像变得更加容易获取，使得该领域的研究加快了步伐。

目前，主要的RGB_D SOD方法可以分为两类：基于传统图像处理技术和基于深度学习的方法（Zhou等，2021）。在基于传统图像处理技术的方法中，通常会利用深度信息来辅助计算图像中的显著区域，例如，基于深度加权的方法、基于深度边缘的方法等。这些方法的优点在于计算效率高，但是对于复杂场景中的显著目标检测效果较差。目前研究主要基于深度学习展开，以提升SOD的性能。这些方法（罗会兰 等，2021）通常采用卷积神经网络（convolutional neural network，CNN）来提取图像特征，并利用深度信息来辅助计算。Fu等人（2020）提出一种用于RGB_D SOD新型联合学习和密集协作融合（joint learning and densely-cooperative fusion，JL-DCF）架构，设计了一个具有良好泛化能力的RGB_D显著目标检测器；Itti等人（1998）提出一种将多级特征重新组合为教师和学生特征的分叉主干策略，引入深度增强模块（depth enhancement module，DEM），能够从通道和空间视图中挖掘深度信息的线索，将RGB和深度模态以互补的方式融合（Sun等，2022）。这些方法在各种数据集上都取得了较好的检测效果。此外，基于图模型的卷积神经网络的方法、基于多任务学习的方法（何静和傅可人，2022）以及基于自注意力机制的图卷积神经网络方法等也引入到RGB_D SOD领域。这些方法在检测性能和计算效率方面都有很大的提升空间。

在探究RGB图像和深度图像融合策略过程中，上述方法均认为两种模态数据在SOD任务上存在互补。本文亦同意该观点，更进一步，认为单一模态数据在SOD任务中无用的信息可能提升另外一种模态数据在SOD任务中的性能。以图1为例，第1行中RGB图像颜色信息是突出目标的主体特征，背景纹理则应是“冗余”特征。但这些对于RGB无用的信息可以辅助对应深度图像过滤由于深度重叠导致的干扰。同理，第2行中由于深度差异较大，深度图像背景的结构信息对前景目标是否能够检出意义不大，但该较为均匀的背景信息对引导RGB图像忽略复杂的背景结构有积极的作用。因此，本文认为在RGB_D SOD任务中两种模态数据通常具有互补的信息存在。

受此启发，提出一种基于互补信息的RGB_D SOD方法。考虑到网络基本卷积结构中常见的线性修正单元（rectified linear unit，ReLU）所具有的选通特点，本文将其进行拓展并用于聚合RGB和深度图像各自的“冗余”信息，并通过监督学习以充分挖掘两种模态数据中的互补信息。

(a) 原始图像　(b)深度图像　(c) 人工标注结果

图1　RGB和深度图像可能存在互补信息的示例图

Fig. 1　An example diagram of where complementary information may exist for RGB and depth images ((a) original images; (b) depth images; (c) manually annotated results)

综上所述，本文的主要贡献如下：1）现有方法虽然构造了多种RGB和深度图的融合策略，但是无法解释两种模态数据中哪些信息有效地构成互补，或者仅能从实验角度证明融合策略的有效性。为此，提出RGB图像和深度图像互补信息聚合模块，该模块能够将两种模态数据各自在常规卷积结构所认为的“冗余”数据进行聚合，显式地形成两种模态数据

互补信息的提取与融合。2)结合互补信息聚合模块,提出了一种RGB_D SOD模型。该模型采用双分支网络框架以提取RGB和深度图像表征特征,并以轻量化网络范式构成主干网络。该网络使用跨模态聚合模块提取两种模态数据的高层语义信息并馈送至主干网络各个尺度特征,最后通过邻域尺度特征增强模块予以合并,以形成特征金字塔结构。实验证明,该种网络范式能够实现高效、准确的显著目标检测性能。3)分析了多种网络监督方法对于提出模型的优化效果,包括深度恢复技术、深度监督技术以及边缘监督技术。并通过消融实验以表明各种监督方法对于网络性能的贡献。4)在4个广泛采用的公共RGB_D数据集上进行定量和定性分析,通过3种常用的评价指标证明了本文提出方法的有效性。

1 相关工作

根据RGB图像和深度数据的互补信息融合方式,现有的RGB_D SOD模型大致分为3类:输入特征级融合模型,输出特征级融合模型和跨模态特征级融合模型(Ji等,2020)。

输入级融合直接连接输入的RGB图像和深度图像,形成一个四通道的输入,然后将其输入到深度卷积神经网络中进行处理,以获得语义信息和几何信息。Qu等人(2017)提出了一种输入级融合模型,该模型首先利用超像素分割获取RGB图像和深度图像的像素集合,并为每个超像素生成特征,再将这些特征输入到卷积神经网络中进行处理以产生每个超像素的显著目标预测。这种输入级融合模型可以充分利用RGB图像和深度数据的互补信息,从而获得准确的检测结果(Liu等,2023)。然而,RGB图像和深度图像表征相差巨大,例如RGB图像的颜色信息经常是目标呈现显著的关键因素,而相应深度图像中的距离信息存在较大的视觉差异性,利用同一网络难以学习各自有效的特征表达。

输出特征级融合方式独立地处理RGB和深度数据,并在网络的末端将两种模态的预测结果进行融合。Wang和Gong(2019)提出了一种输出特征级融合网络(adaptive fusion network, AFNet),该网络可以自适应地融合来自RGB图像和深度图像分支的预测结果,从而获得更好的检测性能。类似地,Piao等人(2019)通过全连接层来融合RGB和深度信息。这种方法可以有效地利用两种模态的特征,从而提高显著目标检测的准确性。此类方法灵活性较高,可以在不同的数据集和任务上进行调整和优化。SOD与分割任务类似,均为密集预测任务,像素语义类别的判定通常需要特征在多个尺度上进行有效交互,然而输出特征级融合方式难以达到该目的。

跨模态(Cong等,2022)融合方法分别提取RGB图像和深度图像特征,通过考虑RGB和深度数据之间的相关性来融合两种模态的中间特征以获得更好的SOD性能 。Chen和Li(2018)提出了一种互补感知融合块,该模块可以有效地融合RGB和深度数据的特征,从而提高目标检测的准确性。何伟和潘晨(2022)设计了一种通道—空间注意力融合模块,利用通道注意力和空间注意力避免融合冗余的背景信息对显著性映射造成影响。Zhang等人(2020)提出了一个互补交互模块,该模块可以从RGB和深度数据中选择互补的表示形式,从而提高SOD的准确性。蒋亭亭等人(2021)设计一种双主干网络和3条解码支路机制,通过特征增强模块实现多模图像的融合互补,从深层次特征中获取全局语义信息,从而得到更加准确的检测结果。Fu等人(2020)提出了一种联合学习和密集协作融合框架,用于互补特征的发现。Piao等人(2020)将深度信息从深度流传递到RGB流,在测试时不使用深度数据,从而实现利用轻量级架构实施SOD任务。这些方法可以提高RGB_D SOD方法的准确性和鲁棒性,并在不同的数据集和任务上获得更好的性能表现。

此外,RGB_D SOD方法已经取得了很好的检测精度(Li等,2024),但其中大部分方法都具有繁重的模型和昂贵的计算成本。这些方法通常需要大量的计算资源和参数,在实际应用中可能会受到限制。因此,为了实现实时检测和轻量级应用,需要进一步研究和开发更加高效和简单的RGB_D显著目标检测方法。可能需要探索一些新的网络架构、损失函数和训练策略,以获得更好的检测性能和更高的计算效率。

2 本文方法

为了独立提取RGB图像和深度图像两种模态数据的深度特征,采用双路主干网络用于提取RGB和深度图像两种模态数据的多尺度特征。由于

RGB图像和深度图像包含不同类型的信息，RGB图像包含颜色和纹理等视觉属性，而深度图像则包含场景中不同物体之间的距离和形状等几何属性，因此，用RGB的视觉属性补充深度图，同时用深度图的几何属性补充RGB图是一种直觉上能够提升性能的合理做法。同时，上述相关工作也大多从该角度出发。然而大规模卷积组，如“卷积+批归一化+激活”构造的信息抽象过程通常非常复杂，难以分析RGB和深度图像中何种信息进行了有效交互。考虑对于输入特征张量x，正负ReLU函数具有简单计算规则，具体为

$$\boldsymbol{F}(x)=\max(0,x)=\begin{cases}x & x\geqslant 0\\ 0 & x<0\end{cases} \tag{1}$$

$$\hat{\boldsymbol{F}}(x)=\min(x,0)=\begin{cases}0 & x\geqslant 0\\ x & x<0\end{cases} \tag{2}$$

式(1)(2)表明，正ReLU函数中只有大于零的卷积响应允许进入下一卷积组，被抑制的卷积响应可以认为是“冗余”的信息。类似地，负ReLU函数则只有小于零的卷积响应允许进入下一卷积组，被抑制的卷积响应认为是“冗余”的信息。以RGB特征为例，其“冗余”信息可能也能够有助于对深度信息形成补充，反之亦可能成立。为此，考虑基于卷积组中激活层对于特征的筛选特性，并以此构造RGB和深度两种模态信息的融合。值得注意的是，Hu和Guo(2021)设计了类似结构以研究自然图像中照度分量和反射分量的分离问题，然而本文讨论的是两种模态数据之间交互问题。此外，本文也是首次将双通道ReLU应用于RGB_D SOD任务。

2.1 整体框架设计

图2展示了本文方法的整体框架。考虑到本文方法针对实时检测场景，采用MobileNetV2(Sandler等，2018)作为提出模型的主干网络。

对于RGB通道，为了使其适应SOD任务，删除了全局平均池化层和最后一个全连接层以减少模型的参数数量和计算复杂度，同时提高特征提取层的重要性和有效性。为了有效融合两种模态数据的互补信息，在主干网络每个降分辨率阶段之后分别添加提出的互补信息聚合模块(complementary-information interaction and aggregation，CIA)模块，CIA是一种交互式双流模块，每个流包含两个注意力机制和一个线性修正单元，这些线性修正单元使用正负ReLU函数，让激活在两个流之间交换，并在被馈送到注意力块之前通过特征级联运算符合并。

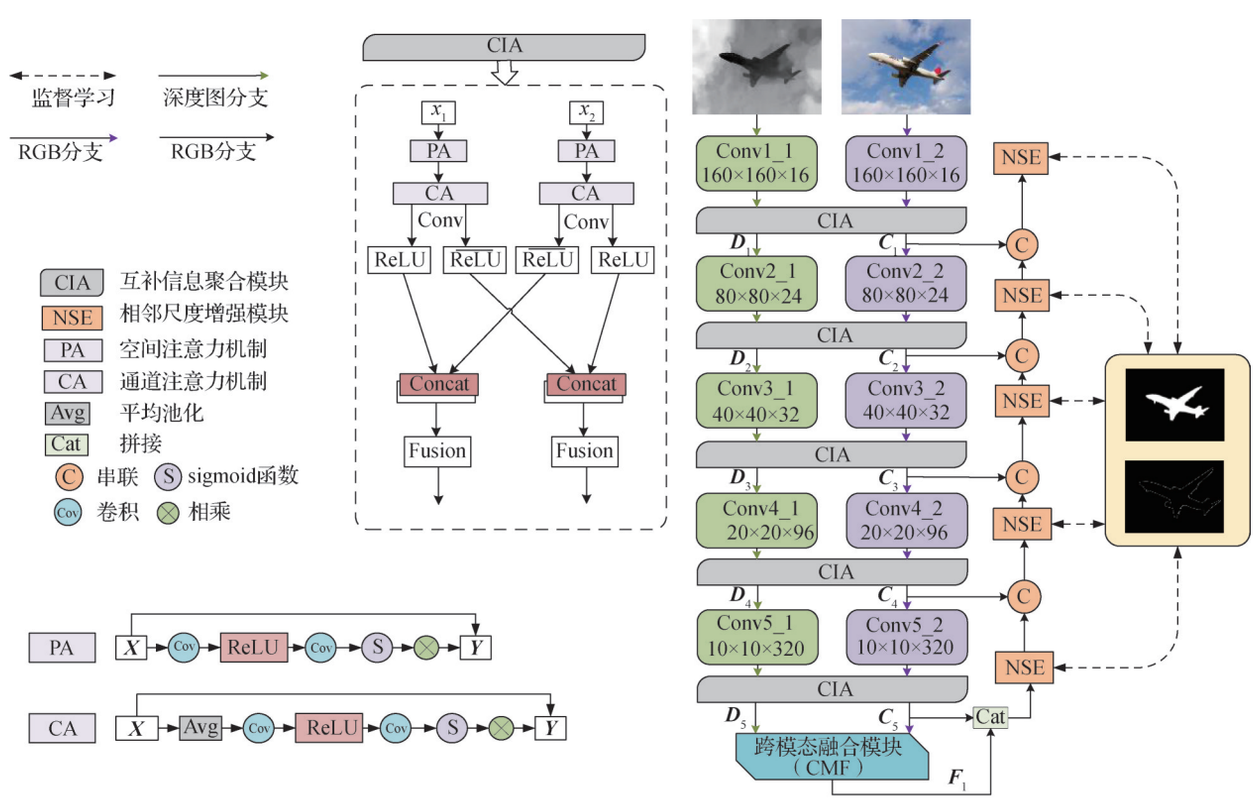

图2 本文方法总体网络结构图

Fig. 2 Diagram of the overall network structure of the proposed method

每个阶段之后将特征图进行2倍率下采样，共进行5次降采样操作。对于网络最顶层RGB和深度特征，本文提出一种跨模态融合模块（cross-modal fusion module，CMF），用于融合两种特征图的语义信息以得到全局特征描述。该全局特征随后反馈至之前各个尺度以形成类似特征金字塔的结构，并通过相邻尺度增强模块（neighboring scale enhancement module，NSE）以逐级向下融合至底层。在训练阶段，本文采用了多种监督方式：1）采用深度恢复监督策略，以保持RGB预测深度信息与深度图像的结构一致性；2）采用深度监督策略，以进一步规范各个尺度特征，使之符合显著目标检测任务的约束；3）采用边缘检测策略，以提高SOD任务中目标的轮廓预测准确性，避免目标边缘模糊。

2.2 互补信息聚合模块（CIA）

CIA旨在利用最简洁的方式提取两种模态数据中的“冗余”信息，并在后续操作中对此类信息进行交互和聚合。其内部包含两个并行的流：首先，在两个支路都设计了两个金字塔注意力模块PA（spatial attention）和CA（channel attention），两者分别通过计算空间注意力矩阵和通道注意力向量以对输入特征进行增强，其结构如图2中对应模块子图所示。在连续进行空间注意力和通道注意力增强后，CIA对两个支路进行信息交互处理，每个流都包含一个卷积层和一个线性修正单元。线性单元使用正负ReLU函数，让激活在两个流之间交换，通过特征级联操作将正ReLU的激活特征和负ReLU的激活特征连接，保证了没有信息从CIA流出，避免了梯度消失等问题，最后再馈送到跨模态融合模块。每个阶段之后本文使用步长为2的卷积层将特征进行降采样，总共进行5次。为方便后续分析，令5个阶段的输出特征图分别为$\boldsymbol{C}_i$，$i=1,2,\cdots,5$。由于深度图具有更少的语义信息，每个阶段只保留两个反向残差块（inverted residual block，IRB）（Sandler等，2018），以使模型更加轻量化。类似地，将5个阶段的输出特征表示为$\boldsymbol{D}_i$，$i=1,2,\cdots,5$。

2.3 跨模态融合模块（CMF）

考虑到语义信息主要存在于RGB图像中，而几何信息主要存在于深度图中，如果能够充分利用两种信息就可以更全面地描述场景和目标。本文提出跨模态融合模块（CMF）用于融合两种特征图的语义信息以得到全局特征描述，提高显著目标检测的准确性，其结构如图3所示。CMF增强了深度特征，并从中获取深度信息，从而更好地引导模型在物体边界处产生更准确的显著目标响应。这种机制可以有效地挖掘深度图像中的语义信息，并提高深度特征的表达能力。首先，本文通过通道注意力机制对深度特征图进行全局平均池化操作，将其转换为一维向量，再通过两个全连接层对该向量进行处理，以提取每个通道的注意力权重。具体计算式为

$$\hat{\boldsymbol{D}}_5=\boldsymbol{D}_5\otimes\sigma(Fc_2(Fc_1(Pool(\boldsymbol{D}_5)))) \tag{3}$$

式中，$Pool(\cdot)$为全局平均池化层，$Fc_1(\cdot)$和$Fc_2(\cdot)$为全连接层，$\sigma(\cdot)$表示sigmoid函数。

加权后的深度特征通过一个3×3卷积层后与RGB特征进行串联，并再次通过两个3×3卷积层以融合两种模态特征，最后逐元素相加得到结合后的特征图，具体过程为

$$\boldsymbol{D}_5^e=\hat{\boldsymbol{D}}_5\oplus Conv(\hat{\boldsymbol{D}}_5) \tag{4}$$

$$\boldsymbol{C}_5^e=\boldsymbol{C}_5\oplus Conv_3\left(Conv_3\left(Cat\left(\boldsymbol{C}_5,\boldsymbol{D}_5^e\right)\right)\right) \tag{5}$$

式中，$\boldsymbol{C}_5$和$\boldsymbol{D}_5$分别是RGB和深度分支的顶层特征，$\boldsymbol{C}_5^e$是融合之后的特征，$Conv_3$表示3×3卷积层。

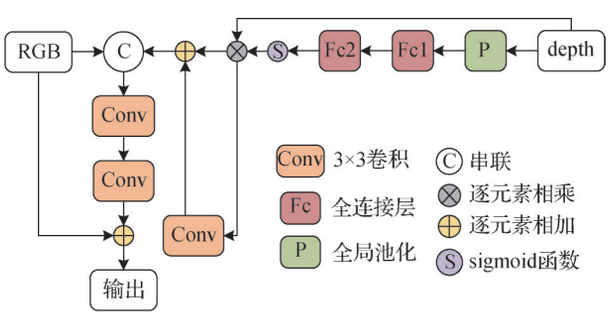

图3 跨模态融合模块网络结构

Fig. 3 Cross-modal fusion module network structure

2.4 相邻尺度增强模块（NSE）

由于底层信息基于像素级别，包含了检测目标的很多细节和位置信息但是缺少全局信息，而高层信息包含关于目标的语义信息，提供了更多的上下文信息（Li等，2020），因此，为了确保在SOD任务中能够综合不同层次的信息，受到Han等人（2018）的启发，本文使用相邻尺度增强模块（NSE）以逐阶段聚合不同尺度特征。NSE采用了空洞卷积ASPP（atrous spatial pyramid pooling）（Chen等，2017）和空间注意力机制的卷积模块，可以增加感受野并提高特征表达能力。如图4所示，为了高效地收集每个空间位置的上下文信息，NSE在原始尺度空间和下

采样空间这两个不同的尺度空间中进行卷积特征变换。特别地，对于输入特征$\boldsymbol{X}$，首先将其沿通道维度均分为两个部分$\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$，设$\boldsymbol{X}_1$为原始尺度空间特征图，其与输入特征具有相同的分辨率用以保持输入特征的细节信息和局部特征。下采样空间$\boldsymbol{X}_2$中则通过下采样降低分辨率以减少计算量，同时扩大感受野，使模型可以捕捉上下文信息。以原始尺度空间特征变换为例，其计算式为

$$\boldsymbol{T}_1 = AvgPool_r(\boldsymbol{X}_1) \tag{6}$$

$$\hat{\boldsymbol{X}}_1 = Up\left(Conv_2(\boldsymbol{T}_1)\right) \tag{7}$$

$$\hat{\boldsymbol{Y}}_1 = Conv_3(\boldsymbol{X}_1) \otimes \sigma(\boldsymbol{X}_1 + \hat{\boldsymbol{X}}_1) \tag{8}$$

式中，$Up(\cdot)$为双线性插值算子，用于将小尺度空间中的中间参考点映射到原始特征空间。σ是sigmoid函数，$\otimes$表示逐元素乘法。最后得到输出，具体为

$$\boldsymbol{Y}_1 = Conv(\hat{\boldsymbol{Y}}_1) \tag{9}$$

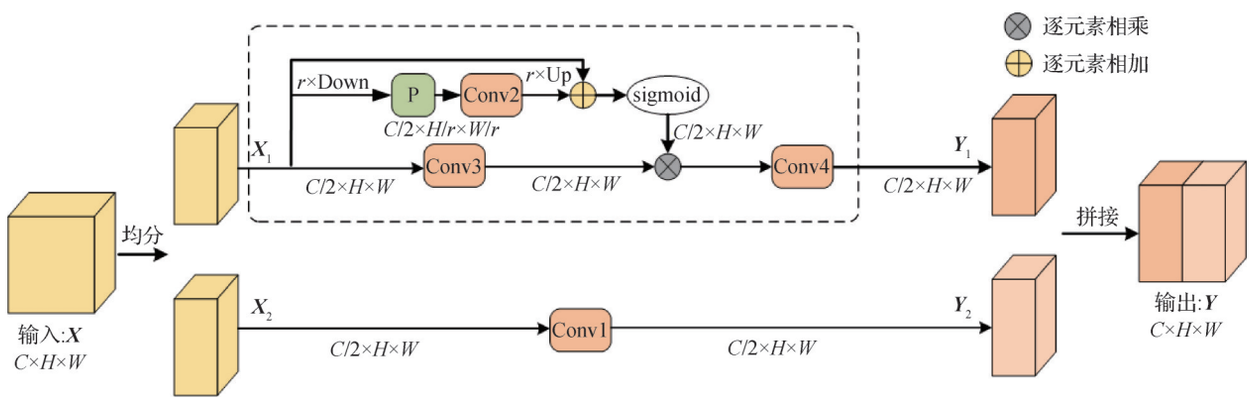

图4　多尺度增强模块网络结构图

Fig. 4　Network structure diagram of multi-scale enhancement module

2.5　网络监督策略

监督学习策略一直是网络优化的关键。为此，构造了多种监督学习策略来更好地导引网络特征的变化。采用的监督学习策略包括以下3种：深度恢复监督、边缘回归监督和深度监督。

2.5.1　深度恢复监督

为了加强RGB_D SOD的特征表示能力，并充分地利用深度图的补充信息，采用深度恢复监督策略来进行网络特征优化，如图5所示。该模块的输入来自主干网络RGB分支4个尺度的输出($\boldsymbol{C}_1$～$\boldsymbol{C}_4$)以及跨模态融合模块CMF的输出$\boldsymbol{F}_1$。设此两类特征拼接结果为$\tilde{\boldsymbol{F}}$，本文采用非局部注意力机制以$\tilde{\boldsymbol{F}}$生成像素级相似度并以此来加权RGB分支底层特征$\boldsymbol{C}_1$以获得恢复的深度图。其计算过程为

$$\boldsymbol{S} = softmax\left(f_v(\tilde{\boldsymbol{F}}) \otimes \left(f_v(\tilde{\boldsymbol{F}})\right)^{\mathrm{T}}\right) \tag{10}$$

$$\boldsymbol{P}_d = \bar{f}_v\left(\boldsymbol{S} \otimes f_v(\boldsymbol{C}_1^*)\right) \tag{11}$$

式中，$f_v(\cdot)$和$\bar{f}_v(\cdot)$为重塑算子，前者将张量由$C \times H \times W$变为$C \times (HW)$；后者则将张量由$C \times (HW)$反向变为$C \times H \times W$。本文采用结构相似性(structure similarity index measure，SSIM)(Wang等，2004)来测量恢复的深度图$\boldsymbol{P}_d$与真实深度图$\boldsymbol{G}_d$之间的结构相似性差异，具体为

$$L_p = 1 - SSIM(\boldsymbol{P}_d, \boldsymbol{G}_d) \tag{12}$$

式中，$SSIM$使用默认参数。值得注意的是，上述非局部注意力机制的计算仅在训练阶段执行，在推理阶段则无需此步骤。因此对网络推理速度没有影响。

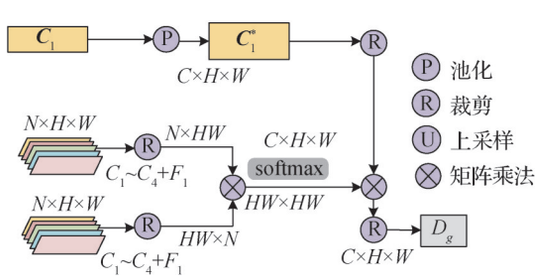

图5　深度恢复模块网络结构图

Fig. 5　Network structure of deep recovery module

2.5.2　边缘监督

考虑到边缘损失能够约束模型对边缘的预测，即使在输入图像中存在一些噪声或者不规则边缘，模型也能正确地分割出物体。使用Canny边缘检测算法来提取图像边缘信息，将这些信息作为额外的监督信号以约束网络预测结果的边缘锐利程度和与

人工标注的一致性。边缘损失采用均方误差(mean squared error,MSE)方式计算,具体为

$$L_{\text{edg}} = MSE(\boldsymbol{P}_e, \boldsymbol{G}_e) \tag{13}$$

式中,$\boldsymbol{P}_e$和$\boldsymbol{G}_e$分别表示预测结果和真实边缘图。

2.5.3 深度监督

采用深度监督策略以进一步规范各个尺度特征,使之符合显著目标检测任务的约束。将网络5个尺度侧方输出的特征进一步进行逐像素预测其对应分辨率的显著图,并与相应缩放的人工标注结果进行损失计算。该部分的损失由BCE(binary crossentropy)损失和Dice(Milletari等,2016)损失组成。设5个尺度的预测结果为$\boldsymbol{P}_i$, $i = 1, 2, \cdots, 5$;对应真值图为$\boldsymbol{G}_t^i$,则每个尺度的损失计算式为

$$L_{\text{sal}}^i = BCE(\boldsymbol{P}_i, \boldsymbol{G}_t^i) + Dice(\boldsymbol{P}_i, \boldsymbol{G}_t^i) \tag{14}$$

式中,BCE表示二进制交叉熵损失函数,其计算式为

$$BCE(\boldsymbol{P}_i, \boldsymbol{G}_t^i) = \boldsymbol{G}_t^i \times \log \boldsymbol{P}_i + (1 - \boldsymbol{G}_t^i) \times \log(1 - \boldsymbol{P}_i) \tag{15}$$

Dice损失的计算式为

$$Dice(\boldsymbol{P}_i, \boldsymbol{G}_t^i) = 1 - \frac{2 \times \boldsymbol{G}_t^i \times \boldsymbol{P}_i}{\|\boldsymbol{G}_t^i\| + \|\boldsymbol{P}_i\|} \tag{16}$$

式中,$\|\cdot\|$表示$\mathcal{L}_1$范数。

综上所述,本文利用3种监督策略以优化网络参数,最终的总损失L的计算式为

$$L = \sum_{i=1}^{5} L_{\text{sal}}^i + \lambda \times L_p + L_{\text{edg}} \tag{17}$$

式中,λ为超参数,用于平衡深度恢复监督的影响,本文取$\lambda = 0.3$。

3 实验结果与分析

3.1 实验设置

本文利用Pytorch(Paszke等,2019)深度学习框架实现所有算法,并在具有GTX3090 GPU和256 GB内存的PC上进行训练,实验中将RGB图像和深度图像统一缩放至320 × 320像素。初始学习率设置为0.000 1,批量大小为16,共训练120个epoch。同时,使用水平翻转和随机裁剪作为默认数据增强,并使用Adam优化器(Kingma和Ba,2016)优化网络。优化器中权重衰减设置为0.000 1,指数衰减率设置为0.99。将提出方法在4个广泛使用的数据集上进行实验,包括NJU2K(Nanjing University 2K)(Ju等,2014)、NLPR(national laboratory of pattern recognition)(Peng等,2014)、STERE(stereo dataset)(Niu等,2012)和SIP(salient person)(Fan等,2020),分别包含1 985、1 200、1 000和927幅图像。在NJU2K、NLPR数据集中随机抽取1 500幅和700幅进行训练,并使用其他485幅NJU2K的图像和300幅NLPR的图像进行测试,另外两个数据集直接用于测试。

3.2 评估标准

采用3个评估指标对模型结果进行评估,其中包括Max F-measure(Achanta等,2009),平均绝对值误差(mean absolute error, MAE),以及Max E-measure(范登平 等,2021)评价指标。

Max F-measure($F_\beta^{\max}$)指在不同的阈值下F-measure(F_β)的最大值。$F_\beta^{\max}$广泛用于衡量二分类算法的性能,其从整体对模型的精准度(precision)和召回率(recall)进行综合评估。精准率和召回率的计算式为

$$P = \frac{TP}{TP + FP} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

式中,TP(true positive)、FP(false positive)和FN(false negative)分别代表真正类、假正类和假反类。精确率和召回率从整体出发,综合评价预测图像的质量,$F_\beta^{\max}$则进一步将这两个指标进行综合,其计算式为

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R} \tag{20}$$

$$F_\beta^{\max} = \max F_\beta \tag{21}$$

式中,β控制P和R的权衡,β^2取值为0.3。$F_\beta^{\max}$越大,表明模型性能越好。

MAE是一种广泛使用的回归模型评价指标,用于衡量预测值与真实值之间的平均差异。其能够避免误差相互抵消的问题,因而可以准确反映实际预测误差的大小。MAE的计算式为

$$f_{\text{MAE}} = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} \left| \boldsymbol{S}(i,j) - \boldsymbol{G}(i,j) \right| \tag{22}$$

式中,$\boldsymbol{S}$和$\boldsymbol{G}$分别表示预测显著图和真值图;(i,j)表示预测显著图的像素位置;w和h分别表示预测显著图的宽和高;$|\cdot|$表示绝对值计算操作。MAE的计算方法简单,其可以累加任何预测结果与人工标注结果之间的差异,因此可以直观地衡量模型的预测能力。MAE值越低,表明模型性能越好。

Max E-measure($E^{\max}$)基于局部像素值和图像平

均值来估计二进制显著图，目的是同时捕捉全局统计量和局部像素匹配信息。其计算式为

$$E = \frac{1}{w \times h}\sum_{i=1}^{w}\sum_{j=1}^{h}\xi(x,y) \tag{23}$$

式中，ξ表示增强对齐矩阵，具体为

$$\xi = \frac{1}{4}(1 + \eta_{FM})^2 \tag{24}$$

由式(24)，得到G和S的偏差矩阵φ_G和φ_S，具体为

$$\eta_{FM} = \frac{2\varphi_G \circ \varphi_S}{\varphi_G \circ \varphi_G + \varphi_S \circ \varphi_S} \tag{25}$$

$$\varphi = I - \mu \times A \tag{26}$$

其中，偏差矩阵φ表示二值映射I与全局平均值μ之间的距离，A是一个元素值均为1的矩阵，大小与I相同。操作符$\circ$表示Hadamard乘积。E^{max}反映了预测值和真值减去它们的全局均值后的相关性。E^{max}值越高，说明模型预测性能越好。

3.3 实验对比

将提出方法与8种具有代表性的方法进行比较，包括CPFP(contrast prior and fluid pyramid integration)(Zhao等，2019)、DMRA(depth-induced multi-scale recurrent attention network)(Piao等，2020)、DANet(dual attention network)(Fu等，2020)、ATSA(asymmetric two-stream architecture network)(Zhang等，2020)、A2dele(adaptive and attentive depth distiller)(Piao等，2020)、D3Net(deep depth-depurator network)(Fu等，2020)、CFID-Net(cascaded feature interaction decoder network)(Chen等，2022)和JSM(joint semantic mining)(Li等，2021)。表1为本文模型与这8种基于深度学习模型在F_β^{max}、MAE、E^{max}评价指标下进行对比的结果。

可以看到，本文方法在4个数据库上的各项指标评估中取得了综合最佳的性能，即最佳或次优结果明显多于其他参与对比的方法。值得注意的是，本文主干网络采用MobileNetV2，使得网络的推理速度达到373.8帧/s，且只有10.8 M个参数。相比于其他方法具有非常显著的速度优势，即本文模型利用较少的参数达到了相当甚至更优的检测性能。

本文方法与其他几种方法的定性评价结果如图6所示。可以看到，本文方法的预测结果在目标内在区域一致性、目标定位的准确性、边缘锐利程度以及与人工标注结果的一致性上均存在优势。

表1 不同方法在4个数据集上Max F-measure、MAE和Max E-measure测度定量评价结果

Table 1 Quantitative evaluation results of Max F-measure, MAE and Max E-measure measures on four datasets of different methods

方法	时间	参数量/M	推力速度/(帧/s)	NJU2K数据集			NLPR数据集			SIP数据集			STERE数据集		
				F_β^{max}	MAE	E^{max}	F_β^{max}	MAE	E^{max}	F_β^{max}	MAE	E^{max}	F_β^{max}	MAE	E^{max}
CPFP	2019	69.5	6	0.850	0.053	0.923	0.840	0.036	0.932	0.821	0.064	0.903	0.889	0.051	0.925
DMRA	2019	59.7	16	0.889	0.051	0.927	0.865	0.031	0.947	0.811	0.086	0.875	0.895	0.047	**0.938**
DANet	2020	26.7	35	0.859	0.053	0.896	0.855	0.035	0.933	0.839	0.063	0.902	0.843	0.054	0.904
ATSA	2020	–	72	**0.896**	0.066	0.879	0.872	0.032	0.948	–	–	–	**0.901**	0.048	0.887
A2dele	2020	–	147	–	–	–	0.871	0.030	0.942	–	–	–	0.879	**0.045**	0.901
D3Net	2021	43.2	78	0.877	0.047	0.913	0.873	0.030	0.944	0.839	0.063	0.902	0.866	0.046	0.920
JSM	2021	–	–	0.674	0.136	0.788	0.743	0.065	0.888	0.622	0.148	0.781	0.748	0.058	0.883
CFID-Net	2022	53.86	39	**0.896**	**0.038**	0.913	**0.892**	0.031	**0.950**	**0.856**	0.057	0.905	0.881	0.047	0.924
本文	2024	**10.8**	**373.8**	**0.896**	0.047	**0.930**	0.874	**0.029**	**0.950**	0.847	**0.054**	**0.907**	0.865	0.046	0.933

注：加粗和下划线字体分别表示最优和次优结果，“–”表示暂无数据。

3.4 消融实验

为了证明本文方法各个模块的有效性，本文针对提出方法中模型结构及监督策略中关键的3个组成部分进行了消融实验，即独立验证互补信息聚合模块、跨模态融合模块、深度恢复监督策略、边缘监督策略。该消融实验采用在NLPR、SIP这两个具有

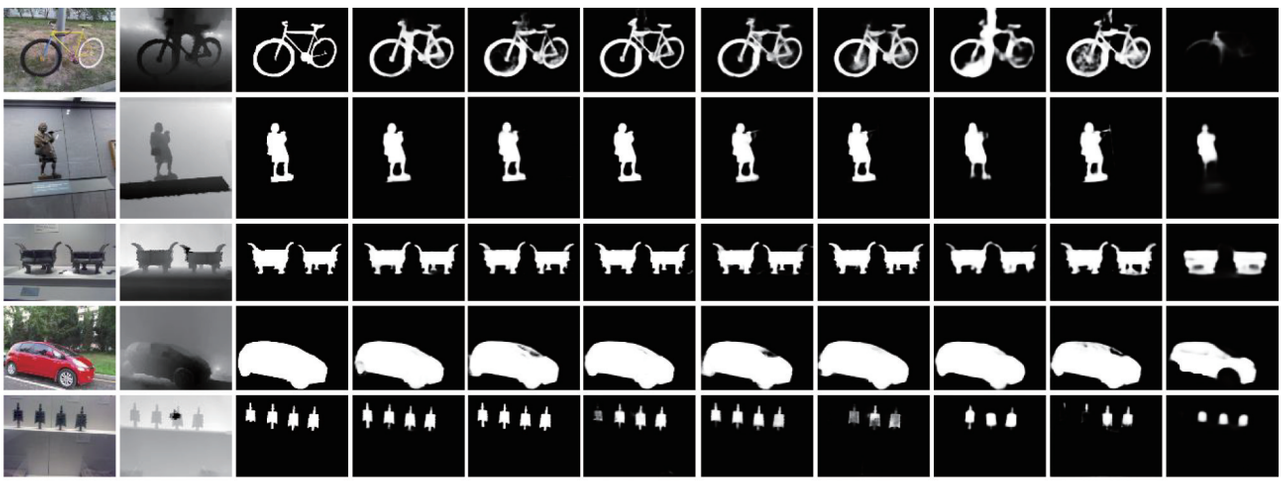

图6　本文方法与其他方法的预测结果的定性评价对比图

Fig. 6　Comparison of qualitative evaluation of prediction results between the proposed method and other methods

代表性的数据集上进行，并使用和横向定量对比中相同的评价指标进行评估。本文在消融实验中遵循和横向对比定量实验中相同的实验设置。实验结果如表2所示。

从表2可以看出，消融任意模块/监督策略均导致最终的性能定量评价产生了下降，该一致性的评价结果充分说明了提出模型的有效性。此外，图7列举了此次消融实验的部分定性比较结果。可以看到，不同的消融策略下网络均不能得到令人满意的视觉效果，即所有评估的消融项目均对本文方法起到了正向的贡献。

值得注意的是，结合表2中结果Ⅳ和结果Ⅵ的定量结果以及图7中的可视化结果，可以进一步分析出互补信息聚合模块（CIA）对于网络整体性能的提升。当移除CIA后可发现定量和定性结果均明显有性能下降，表明本文提出的两种模态互补信息聚合的思路能够有效地提升RGB_D SOD任务的性能。此外，图2中所示的CIA模块具有较为简单的结构，因此可以方便地应用于具有双分支网络的RGB_D SOD模型中。

表2　本文方法中模型结构及监督策略的关键组件的定量消融实验结果

Table 2　Quantitative ablation results of key components of the model structure and surveillance strategy in proposed method

结果	方法				NLPR			SIP		
	CIA	CMF	深度监督	边缘损失	$F_\beta^{\max}$	MAE	$E^{\max}$	$F_\beta^{\max}$	MAE	$E^{\max}$
Ⅰ	√	√	–	–	0.845	0.033	0.878	0.835	0.071	0.863
Ⅱ	√	√	√	–	0.826	0.038	0.876	0.813	0.078	0.855
Ⅲ	√	–	√	√	0.822	0.039	0.877	0.827	0.076	0.867
Ⅳ	–	√	√	√	0.843	0.047	0.865	0.825	0.088	0.825
Ⅴ	√	√	–	√	0.866	0.038	0.851	0.836	0.079	0.866
Ⅵ	√	√	√	√	**0.874**	**0.030**	**0.907**	**0.845**	**0.060**	**0.907**

注：加粗字体表示各列最优结果，“√”表示采用，“–”表示未采用。

在本文采用的监督策略中深度恢复监督损失由超参数λ决定其对于网络参数调整的影响。消融实验中针对不同λ的取值进行了定量的评价，结果如表3。可以看出，当$\lambda = 0.3$时，网络整体性能达到最佳，因此采用$\lambda = 0.3$作为所有实验的默认设置。

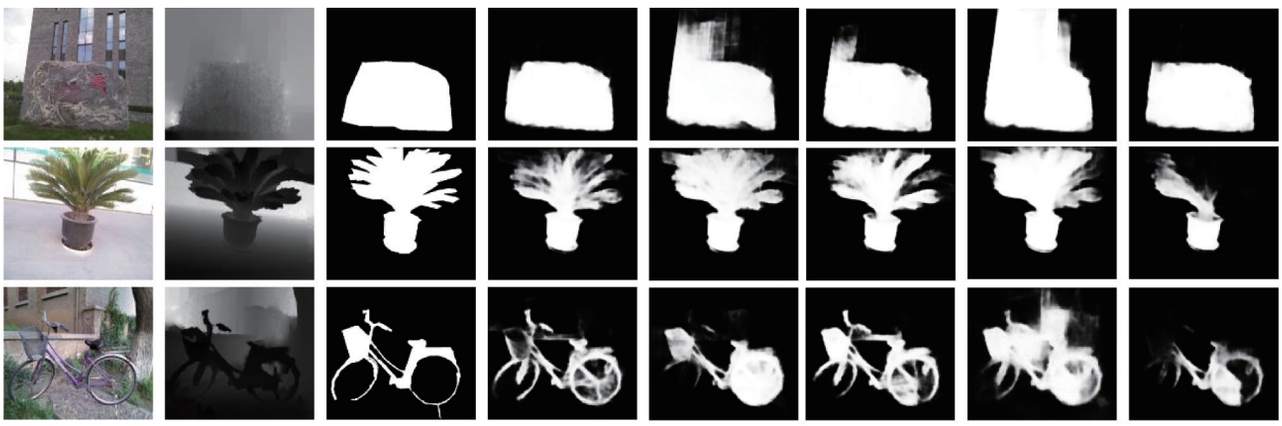

图7　本文方法中模型结构及监督策略的关键组件的定性消融实验结果

Fig. 7　Qualitative ablation results of key components of the model structure and monitoring strategy in the proposet method

表3　深度恢复监督损失权重λ系数对于整体性能影响的定量评价结果

Table 3　Quantitative evaluation results of the influence of depth recovery supervision loss weight λ coefficient on the overall performance

λ	F_β	MAE
0	0.834	0.056
0.1	0.825	0.055
0.2	0.865	0.048
0.3	**0.875**	**0.047**
0.4	0.844	0.047
0.5	0.821	0.049
0.6	0.754	0.056
1	0.842	0.047

注：加粗字体表示各列最优结果。

4　结　论

针对RGB_D SOD任务，首先回顾了在RGB_D SOD任务中融合RGB图像和深度图像信息方面的现有工作，并总结了合理利用两种模态数据互补信息对于性能提升的重要性。在探讨了现有方法在该方面存在的问题之后，本文分析了现有网络基本卷积组结构中线性修正单元的选通行为，并进一步提出了一种互补信息交互融合（CIA）模块，将单一模态的“冗余”特征用于辅助另一模态特征。以此思路为基础，提出了一种新的RGB_D SOD模型，该模型结合轻量级主干网络与设计的跨模态特征融合（CMF）模块、邻域尺度特征增强（NSE）模块等一起构成了多尺度特征金字塔结构框架。为了有效监督提出模型的优化过程，采用了3种监督策略，包括深度恢复监督、边缘监督和深度监督。在4个广泛使用的公开数据集上的定量和定性的实验结果表明，以3种主流测度作为评估方法，提出方法相比较参与对比的方法取得了更优秀的性能。同时，本文通过消融实验进一步分析了本文方法的网络结构和监督策略中关键部分对于网络整体性能的贡献。定量和定性的消融实验结果表明这些关键组件及策略对于网络性能均起到正向的作用。但这些公开数据库中的多数样本场景复杂度有限，未考虑到真实环境中的不利条件，如光线变化、玻璃反射等。因此，未来研究的一个关键方向是将所提算法应用于真实场景，通过不断改善更全面地评估其性能以及实际应用价值。

参考文献（References）

Achanta R, Hemami S, Estrada F and Susstrunk S. 2009. Frequency-tuned salient region detection//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009: 1597-1604 [DOI: 10.1109/CVPR.2009.5206596]

Chen H and Li Y F. 2018. Progressively complementarity-aware fusion network for RGB-D salient object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3051-3060 [DOI: 10.1109/CVPR.2018.00322]

Chen L C, Papandreou G, Schroff F and Adam H. 2017. Rethinking

atrous convolution for semantic image segmentation [EB/OL]. [2023-08-23]. https://arxiv.org/pdf/1706.05587.pdf

Chen T Y, Hu X G, Xiao J, Zhang G F and Wang S J. 2022. CFIDNet: cascaded feature interaction decoder for RGB-D salient object detection. Neural Computing and Applications, 34(10): 7547-7563 [DOI: 10.1007/s00521-021-06845-3]

Cong R M, Lin Q W, Zhang C, Li C Y, Cao X C, Huang Q M and Zhao Y. 2022. CIR-Net: cross-modality interaction and refinement for RGB-D salient object detection. IEEE Transactions on Image Processing, 31: 6800-6815 [DOI: 10.1109/TIP.2022.3216198]

Cong R M, Zhang C, Xu M, Liu H Y and Zhao Y. 2023. Research progress of RGB-D salient object detection in deep learning era. Journal of Software, 34(4): 1711-1731 (丛润民, 张晨, 徐迈, 刘鸿羽, 赵耀. 2023. 深度学习时代下的RGB-D显著性目标检测研究进展. 软件学报, 34(4): 1711-1731) [DOI: 10.13328/j.cnki.jos.006700]

Fan D P, Ji G P, Qin X B and Cheng M M. 2021. Cognitive vision inspired object segmentation metric and loss function. SCIENTIA SINICA Informationis, 51(9): 1475-1489 (范登平, 季葛鹏, 秦雪彬, 程明明. 2021. 认知规律启发的物体分割评价标准及损失函数. 中国科学: 信息科学), 51(9): 1475-1489 [DOI: 10.1360/SSI-2020-0370]

Fan D P, Lin Z, Zhang Z, Lin Z, Zhu M L and Cheng M M. 2020. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems, 32(5): 2075-2089 [DOI: 10.1109/TNNLS.2020.2996406]

Fu K R, Fan D P, Ji G P and Zhao Q J. 2020. JL-DCF: joint learning and densely-cooperative fusion framework for RGB-D salient object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3049-3059 [DOI: 10.1109/CVPR42600.2020.00312]

Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W and Lu H Q. 2019. Dual attention network for scene segmentation//Proceedings of 2019 IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA: IEEE: 3146-3154 [DOI: 10.1109/CVPR.2019.00326]

Han J W, Chen H, Liu N, Yan C G and Li X L. 2018. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics, 48(11): 3171-3183 [DOI: 10.1109/TCYB.2017.2761775]

He J, Fu K R. 2022. RGB-D salient object detection of using few-shot learning. Journal of Image and Graphics, 27(10): 2860-2872 (何静, 傅可人. 2022. 小样本条件下的RGB-D显著性物体检测. 中国图象图形学报, 27(10): 2860-2872) [DOI: 10.11834/jig.211068]

He W and Pan C. 2022. The salient object detection based on attention-guided network. Journal of Image and Graphics, 27(4): 1176-1190 (何伟, 潘晨. 2022. 注意力引导网络的显著性目标检测. 中国图象图形学报, 27(4): 1176-1190) [DOI: 10.11834/jig.200658]

Hu Q M and Guo X J. 2021. Trash or treasure? An interactive dual-stream strategy for single image reflection separation//Proceedings of the 35th Conference on Neural Information Processing Systems. [s.l.]: NeurIPS: 24683-24694

Itti L, Koch C and Niebur E. 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11): 1254-1259 [DOI: 10.1109/34.730558]

Ji W, Li J J, Zhang M, Piao Y R and Lu H C. 2020. Accurate RGB-D salient object detection via collaborative learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 52-69 [DOI: 10.1007/978-3-030-58523-5_4]

Jiang T T, Liu Y, Ma X and Sun J L. 2021. Multi-path collaborative salient object detection based on RGB-T images. Journal of Image and Graphics, 26(10): 2388-2399 (蒋亭亭, 刘昱, 马欣, 孙景林. 2021. 多支路协同的RGB-T图像显著性目标检测. 中国图象图形学报, 26(10): 2388-2399) [DOI: 10.11834/jig.200317]

Ju R, Ge L, Geng W J, Ren T W and Wu G S. 2014. Depth saliency based on anisotropic center-surround difference//Proceedings of 2014 IEEE International Conference on Image Processing. Paris, France: IEEE: 1115-1119 [DOI: 10.1109/ICIP.2014.7025222]

Kingma D P and Ba J. 2016. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-15

Li C Y, Cong R M, Piao Y R, Xu Q Q and Loy C C. 2020. RGB-D salient object detection with cross-modality modulation and selection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 225-241 [DOI: 10.1007/978-3-030-58598-3_14]

Li J J, Ji W, Bi Q, Yan C, Zhang M, Piao Y R, Lu H C and Cheng L. 2021. Joint semantic mining for weakly supervised RGB-D salient object detection//Proceedings of the 35th Conference on Neural Information Processing Syste. [s.l.]: NeurIPS: 11945-11959

Li L, Han J W, Liu N, Khan S, Cholakkal H, Anwer R M and Khan F S. 2024. Robust perception and precise segmentation for scribble-supervised RGB-D saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(1): 479-496 [DOI: 10.1109/TPAMI.2023.3324807]

Liu C, Yang G, Wang S, Wang H X, Zhang Y H and Wang Y T. 2023. TANet: transformer-based asymmetric network for RGB-D salient object detection. IET Computer Vision, 17(4): 415-430 [DOI: 10.1049/cvi2.12177]

Luo H L, Yuan P and Tong K. 2021. Review of the methods for salient object detection based on deep learning. Acta Electronica Sinica, 49(7): 1417-1427 (罗会兰, 袁璞, 童康. 2021. 基于深度学习的显著性目标检测方法综述. 电子学报, 49(7): 1417-1427) [DOI: 10.12263/DZXB.20200651]

Milletari F, Navab N and Ahmadi S A. 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation//Proceedings of the 4th International Conference on 3D Vision. Stanford, USA: IEEE: 565-571 [DOI: 10.1109/3DV.2016.79]

Niu Y Z, Geng Y J, Li X Q and Liu F. 2012. Leveraging stereopsis for saliency analysis//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 454-461[DOI: 10.1109/CVPR.2012.6247708]

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z M, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Lu F, Bai J J and Chintala S. 2019. PyTorch: an imperative style, high-performance deep learning library//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: 1-12

Peng H W, Li B, Xiong W H, Hu W M and Ji R R. 2014. RGBD salient object detection: a benchmark and algorithms//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 92-109[DOI: 10.1007/978-3-319-10578-9_7]

Piao Y R, Ji W, Li J J, Zhang M, and Lu H C. 2019. Depth-induced multi-scale recurrent attention network for saliency detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7253-7262 [DOI: 10.1109/ICCV.2019.00735]

Piao Y R, Rong Z K, Zhang M, Ren W S and Lu H C. 2020. A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9057-9066[DOI: 10.1109/CVPR42600.2020.00908]

Qu L Q, He S F, Zhang J W, Tian J D, Tang Y D and Yang Q X. 2017. RGBD salient object detection via deep fusion. IEEE Transactions on Image Processing, 26(5): 2274-2285 [DOI: 10.1109/TIP.2017.2682981]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Sun H, Liu Y S and Lin Y H. 2023. Deep learning based salient object detection: a survey. Journal of Data Acquisition and Processing, 38(1): 21-50 (孙涵, 刘译善, 林昱涵. 2023. 基于深度学习的显著性目标检测综述. 数据采集与处理, 38(1): 21-50)[DOI: 10.16337/j.1004-9037.2023.01.002]

Sun P, Zhang W H, Li S Y, Guo Y L, Song C L and Li X. 2022. Learnable depth-sensitive attention for deep RGB-D saliency detection with multi-modal fusion architecture search. International Journal of Computer Vision, 130(11): 2822-2841 [DOI: 10.1007/s11263-022-01646-0]

Wang N N and Gong X J. 2019. Adaptive fusion for RGB-D salient object detection. IEEE Access, 7: 55277-55284 [DOI: 10.1109/ACCESS.2019.2913107]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]

Zhang J, Fan D P, Dai Y C, Anwar S, Saleh F S, Zhang T and Barnes N. 2020. UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8579-8588 [DOI: 10.1109/CVPR42600.2020.00861]

Zhang M, Fei S X, Liu J, Xu S, Piao Y R and Lu H C. 2020. Asymmetric two-stream architecture for accurate RGB-D saliency detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 374-390 [DOI: 10.1007/978-3-030-58604-1_23]

Zhao J X, Cao Y, Fan D P, Cheng M M, Li X Y and Zhang L. 2019. Contrast prior and fluid pyramid integration for RGBD salient object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3922-3931 [DOI: 10.1109/CVPR.2019.00405]

Zhou T, Fan D P, Cheng M M, Shen J B and Shao L. 2021. RGB-D salient object detection: a survey. Computational Visual Media, 7(1): 37-69 [DOI: 10.1007/s41095-020-0199-z]

作者简介

叶欣悦，女，硕士研究生，主要研究方向为目标检测与识别、深度学习和语义分割。E-mail: 1563228498@qq.com

朱磊，通信作者，男，副教授，主要研究方向为目标检测与识别、语义分割和场景解析。E-mail: zhulei@wust.edu.cn

王文武，男，副教授，主要研究方向为目标检测与识别、语义分割和场景解析。E-mail: 46276773@qq.com

付云，男，硕士研究生，主要研究方向为计算机视觉、深度学习、语义分割和弱监督语义分割。

E-mail: 1666314753@qq.com