〈综述与评论〉

RGB-T 显著性目标检测综述

吴锦涛，王安志，任春洪
（贵州师范大学 大数据与计算机科学学院，贵州 贵阳 550000）

摘要：除 RGB 图像外，热红外图像也能提取出对显著性目标检测至关重要的显著性信息。热红外图像随着红外传感设备的发展和普及已经变得易于获取，RGB-T 显著性目标检测已成为了热门研究领域，但目前仍缺少对现有方法全面的综述。首先介绍了基于机器学习的 RGB-T 显著性目标检测方法，然后着重介绍了两类基于深度学习的 RGB-T 显著性目标检测方法：基于卷积神经网络和基于 Vision Transformer 的方法。随后对相关数据集和评价指标进行介绍，并在这些数据集上对代表性的方法进行了定性和定量的比较分析。最后对 RGB-T 显著性目标检测面临的挑战及未来的发展方向进行了总结与展望。

关键词：显著性目标检测；热红外图像；RGB-T 显著性目标检测；深度学习
中图分类号：TP391 **文献标识码**：A **文章编号**：1001-8891(2025)01-0001-09

RGB-T Salient Object Detection: A Survey

WU Jintao，WANG Anzhi，REN Chunhong
(*School of Big Data and Computer Science, Guizhou Normal University, Guiyang* 550000, *China*)

Abstract: In addition to RGB images, thermal IR images can be used to extract salient information, which is crucial for salient object detection. With the development and popularization of IR sensing equipment, thermal IR images have become readily available, and RGB-T salient object detection has become a popular research topic. However, there is currently a lack of comprehensive surveys on the existing methods. First, we briefly introduce machine learning-based RGB-T salient object detection methods and then focus on two types of deep learning methods based on CNNs and vision transformers. Subsequently, relevant datasets and evaluation metrics are introduced, and both qualitative and quantitative comparative analyses are conducted on representative methods using these datasets. Finally, challenges and future development directions for RGB-T salient object detection are summarized and discussed.

Key words：salient object detection, infrared image, RGB-T salient object detection, deep learning

0 引言

人类的视觉感知系统能够捕获场景中的重要物体和场景信息，如颜色、轮廓、景深等属性，通过处理和整合这些信息，人类可以在各种场景下迅速定位感兴趣的区域，即显著性区域。显著性目标检测就是模拟上述过程，赋予计算机系统快速定位重要目标、感知场景重要信息的能力，已经被广泛应用于图像分类[1]、语义分割[2]和目标识别[3]等众多计算机视觉任务。

显著性目标检测最早在 1998 年被 Itti 等人[4]提出。此后一段时间内，显著性目标检测虽然得到了一定的发展，但面对复杂背景、光照变化等挑战性因素难以取得理想效果。近年来，随着热红外传感器的发展和普及，一些研究发现热红外信息在处理照明条件、复杂背景等因素导致的目标模糊问题时具有很好的效果[5-6]，非常适合处理低光、雨雾天气等恶劣条件拍摄的图像。为了提升显著性目标检测的性能，一些学者以 RGB 和热力信息为输入，设计

收稿日期：2023-11-01；**修订日期**：2024-01-19.
作者简介：吴锦涛（2000-），男，浙江宁波人，硕士研究生，研究方向：显著性目标检测。E-mail：bigdatawujitnao@163.com。
通信作者：王安志（1986-），男，贵州铜仁人，副教授，研究方向：人工智能、计算机视觉等。E-mail：cvmll6102@163.com。
基金项目：国家自然科学基金地区基金项目（62162013）；贵州师范大学学术新苗基金项目（黔师新苗[2022]30 号）。

了 RGB-T 显著性目标检测方法[7-8]。2017 年 Ma 等人[8]提出了首个 RGB-T 显著性目标检测方法（Multiscale Features and SVM Regressors，MFSR），该方法利用 VGG16 分别生成 RGB 和热力图像的显著预测，然后训练一个支持向量机来有权重地融合两个显著图，验证了热力信息能够有效提升检测的效果。自此开始，一系列 RGB-T 显著性目标检测方法被先后提出[9-11]，但目前仍没有相关的综述对现有的方法进行梳理。鉴于此，本文首次对 RGB-T 显著性目标检测进行系统全面的综述，旨在总结这些方法、梳理脉络、了解其当前发展趋势、探明未来发展方向与研究重点，为 RGB-T 显著性目标检测的发展提供参考。

如图 1 所示，RGB-T 显著性目标检测可分为基于机器学习和基于深度学习两类方法。以 Wang 等人[7]提出的首个方法 MTMR（Multi-task Manifold Ranking）为例，基于机器学习的 RGB-T 显著性目标检测方法存在方法单一、鲁棒性差的问题，面对复杂情况往往效果较差甚至失效。以 Ma 等人[8]提出的首个方法 MFSR 为例，基于深度学习的 RGB-T 显著性方法具有端到端学习、可扩展性强等诸多优点，非常适合处理大规模复杂数据，目前已成为 RGB-T 显著性目标检测中的主流，也是本文重点阐述的内容。图 2 展示了 RGB-T 显著性目标检测发展时间线。

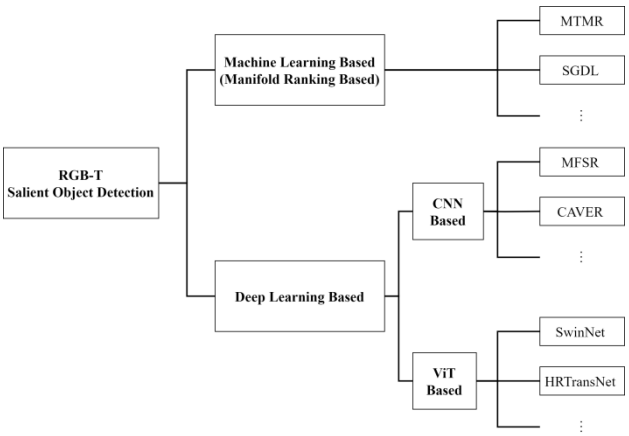


图 1 RGB-T 显著性目标检测的分类

Fig.1 Classification of RGB-T salient object detection

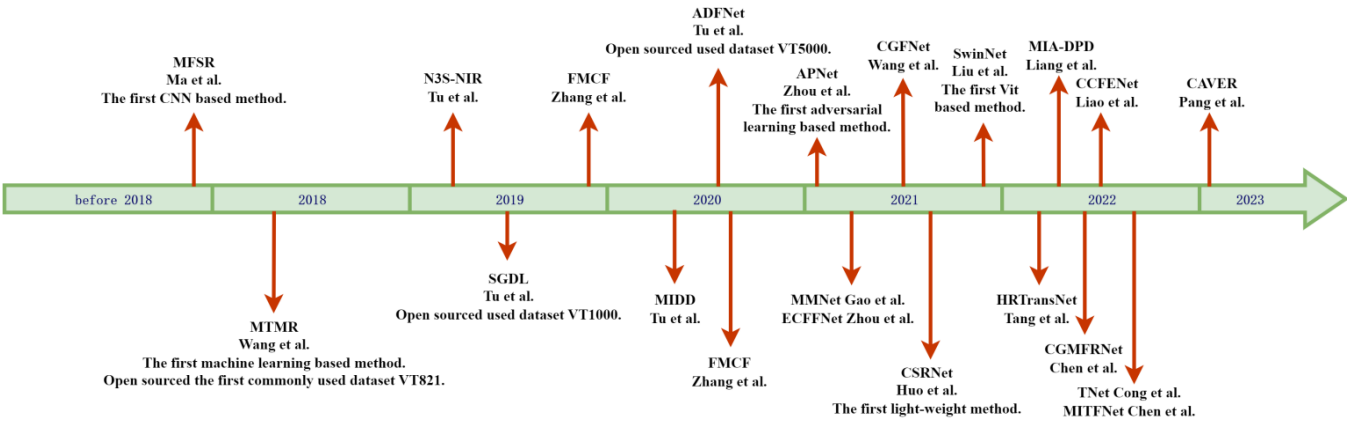


图 2 RGB-T 显著性目标检测的发展历程

图 2 The development of RGB-T saliency target detection

1 RGB-T 显著性目标检测方法

1.1 基于机器学习的 RGB-T 显著性目标检测

现有基于机器学习的 RGB-T 显著性目标检测都采用图流形排序（Graph-based Mainfold Ranking，GMR）算法。2004 年，Zhou 等人[9]首次提出了流形排序的概念。简单来说，流形排序以传播和迭代的方式计算出数据集中数据点成为“中心类”的可能性，而 GMR 以图模型构建数据点进行流形排序。2018 年，Wang 等人[7]首次将 GMR 引入 RGB-T 显著性目标检测中，提出了 MTMR 算法。MTMR 将热力信息视为 RGB 图像额外的通道，利用图像的超像素作为初始节点，以边界节点作为背景类进行流形排序定位显著区域，再以显著区域中的节点作为显著中心进行流形排序，并根据模态可靠性和一致性自适应地融合双模态信息。此外，他们还提出了首个被广泛使用的 RGB-T 显著性目标检测数据集 VT821[7]，为后续的研究奠定了基础。

在 GMR 算法中，节点的选择和排序算法会对预测结果起重要影响。Tu 等人[10]针对边界噪声和真实场景中的目标存在大小或内外部不一致的情况提出一个多尺度噪声不敏感的方法。该方法利用多粒度超像素分割获取多尺度图节点，并通过引入中间变量优化了节点选择的过程，从而降低了显著性结果的边界噪声，且面对处于边界的目标取得了较好的效果。Huang 等人[11]设计了一个低秩亲和矩阵来建模超像素关系，并在流形排序之前根据亲和矩阵

和学习重要特征,结合多模态一致性和异构性自适应地融合多模态特征。Huang 等人[12]基于双模态的边界信息捕获了更精确的超像素,并提出了多图融合模型来选择性地从多模态图像中学习有用信息。

简言之,此类方法均需要通过人为设计提取出有用的特征,严重依赖设计者的经验和水平,设计者往往需要耗费大量的人力和时间来进行特征设计和调优,在面对复杂场景下的检测效果不佳,存在目标模糊、检测错误、检测缺失等问题。

1.2 基于深度学习的 RGB-T 显著性目标检测

相较于机器学习,深度学习在自动提取特征、处理复杂数据、可扩展性、端到端学习、模型容量等方面均更具优势,在计算机视觉领域获得了巨大成功[13-14],也被引入 RGB-T 显著性目标检测中,尤其以卷积神经网络(Convolutional Neural Networks,CNN)和 Vision Transformer(ViT)的表现优异。本章将分别介绍基于 CNN 和基于 ViT 的 RGB-T 显著性目标检测方法。

1.2.1 基于 CNN 的 RGB-T 显著性目标检测

CNN 的特点是对图像的局部信息比较敏感,模型结构相对简单、参数较少,非常适合显著性目标检测任务。2017 年,Ma 等人[7]提出了第一个基于 CNN 的 RGB-T 显著性目标检测的方法 MFSR,随后出现了引入注意力机制、探索上下文信息等基于不同设计的方法。本文将这些方法进一步分为基于超像素的方法、基于特征增强的方法和基于特征融合的方法 3 类。

1)基于超像素的方法。超像素是基于机器学习的 RGB-T 显著性目标检测中最常用的结构。为了克服此类方法通过人为设计提取特征的局限,部分学者引入 CNN 特征探索超像素之间的关系。Tu 等人[15]使用 VGG19 提取特征,在每个模态的多层次特征上进行流形排序,并利用亲和力矩阵提出了一种联合图学习(Saliency Detection via Collaborative Graph Learning,SDCGL)方法。Pang 等人[16]基于背景的加权图来初步定位深度特征的显著性区域,再利用设计的跨模态协同反馈细胞自动机(Cross-modal Co-feedback Cellular Automata)细化显著目标。不同于上述方法利用深度特征建模超像素关系,Liu 等人[17]提出了一个基于涂鸦的弱监督框架,通过涂鸦区域的超像素来定位目标,约束双模态深层特征分别生成显著性预测,预测图被聚合为伪标签用于监督目标网络的显著性预测结果。

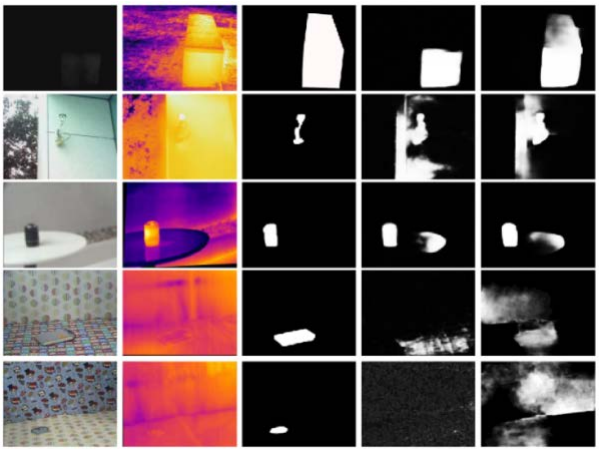
尽管上述方法利用深度特征克服了手工特征(hand-crafted feature)的局限性,但是仍然存在检测精度不理想、边界粗糙等问题,其检测结果也呈现明显的超像素纹理。

2)基于特征增强的方法。特征增强是一种常用的处理特征数据的方法,通过对特征数据进行一系列操作和变换得到更具判别性的特征,以提升模型的特征表达能力。捕获 RGB 和热力图像中的有效信息是 RGB-T 显著性目标检测的一个关键问题,也是基于特征增强的方法主要解决的问题。Zhang 等人[18]最先意识到了该问题,提出了一个近层特征融合(Fusing Multi-level CNN Features,FMCF)的方法,它将每个模态的中间三层特征分别与各自的相邻层特征聚合,探索了多层次特征的互补性。Zhang 等人[19]通过级联不同膨胀率的空洞卷积层和最大池化层设计了一个混合空洞池化模块,探索了多尺度下特征的上下文信息,在丰富了特征尺度的同时,增强了全局和局部特征的空间一致性。Bi 等人[20]设计了一个并行对称融合模块,以并行对称的方式整合相邻层的关键显著性线索进行显著性预测。Tu[21]引入了基于卷积的注意力模块,该模块级联了通道注意力和空间注意力,能够过滤掉特征中不重要的信息。Tu[22]等人和 Wang 等人[23]挖掘网络最深层语义信息来引导浅层特征获取更加有效的特征表示。

图 3 展示了经典的 RGB-D SOD 方法 RD3D(RGB-D SOD via 3D CNN)[24]和早期具有代表性的 RGB-T SOD 方法 MIDD(Multi-Interactive Dual-Decoder)[22]在低质量 RGB 图像(第 1 行)、低质量热红外图像(第 2 行)、高质量 RGB-T 图像(第 3 行)和复杂背景(第 4、5 行)下的结果。从图中可以看出,RGB 和热红外图像的成像质量会受各种因素(如温度、光照等)的影响,不加以区分地使用它们会导致网络受到严重的噪声干扰。此外,RGB 和热成像两种模态信息是从不同角度表达画面的特征,是同一张图像的不同属性,在特征表示层面存在差异,不当地处理差异信息也同样会引入噪声。因此,一些方法考虑如何有效地探索多模态特征之间一致性与差异性来补充或过滤低质量的单模态信息。Chen 等人[25]设计了一个特征差异缩减模块,通过捕获的差异性信息增强双模态特征的一致性。Liao 等人[26]提出一个交叉协同增强策略,自适应地从每个模态中收集更有效的特征表示,并协同纠正有缺陷的特征响应。与上述方法不同,Cong 等人[27]认为低光场景下的 RGB 图像质量不佳,因此引入了光照度评分的方法评估图像质量,避免了低质量的输入。

3)基于特征融合的方法。作为多模态视觉任务,如何有效地融合多模态特征是 RGB-T 显著性目标检测的另一个关键问题,基于特征融合的方法旨在充分挖掘两种模态间的互补属性并充分利用它们各

自的优势。一些方法直接地融合多模态特征[17-18]，但这种方式容易受到噪声信息和冗余数据的限制。现有的方法主要考虑如何自适应地、有选择地融合两个模态的特征，例如：Liang 等人[28]提出了一个多模态交互注意单元，它可以学习内容相关的权重向量自适应地融合重要的多模态信息。Chen 等人[25]设计了一个基于注意力的交叉注意融合模块，分别在通道和空间层面融合更加重要的特征。Gao 等人[29]受到视觉色相学说中视觉颜色感知的有效选择和对抗过程的启发，设计了多阶段的特征融合方法 MMNet（Multi-stage and Multi-scale fusion Network）。该方法将有效特征的捕获和融合对应上述两个过程，可以有效地增强显著区域并抑制多模态的不一致性和低质量数据的影响。Pang 等人[30]组合自注意力和交叉注意力提出了一个跨模态聚合单元，根据语义相似性从全局序列中的其他特征收集信息，并为两种注意力计算增加了通道分支，充分利用空间和通道视角下的模态间和模态内信息。此外，考虑到两种注意力的计算复杂度，他们还设计了一个无参数逐像素的补丁嵌入来降低运算开销。



(a) RGB (b)Thermal (c) GT (d) RD3D (e)MIDD

图 3 不同质量的输入及其显著性预测

Fig.3 Different quality inputs and their salient prediction

总的来说，从上述方法我们可知上下文语义、模态差异等信息对 RGB-T 显著性目标检测任务至关重要，而注意力机制能有效减少噪声干扰，特别是自注意力能够建立全局上下文信息，能够在一定程度上克服 CNN 不擅长捕获全局特征的问题。除此之外，还有一些方法考虑到边缘设备有限的计算资源，设计了轻量化的 RGB-T 显著性目标检测方法[31-32]。尽管基于 CNN 的 RGB-T 显著性目标检测方法受到了广泛的关注，由于 CNN 本身存在更关注局部特征，对位置信息不敏感，这些方法在面对一些需要感知全局语义的场景仍然存在不足，还需要进一步的探索和研究。

1.2.2 基于 ViT 的 RGB-T 显著性目标检测

得益于强大的全局特征建模能力，ViT 克服了 CNN 关注局部信息的局限性，在其被引入计算机视觉领域后，基于 ViT 的方法在多个任务中拥有优异的表现[33-34]。ViT 将输入图像转化为一个序列，并为其添加位置编码以保证处理过程中的序列关系。自注意力机制是 ViT 的核心组件，通过学习，它可以为输入建立 Q（query）、K（key）和 V（value）3 种映射表示，从而计算序列中每个位置的相关性。通过调整 Q、K、V 的使用，ViT 可以获取全局上下文关系，构建丰富的信息表示。2021 年，首个基于 ViT 的 RGB-T 显著性目标检测方法 SwinNet[35]（Swin Transformer drives Network）被提出，显示出此类方法巨大的发展潜力。随后也有一些研究对此类方法进行了探索，但总体上仍处于发展的初级阶段，本节将逐一对现有基于 ViT 的 RGB-T 显著性目标检测方法进行介绍。

Liu 等人[35]基于 Swin-Transformer 提出了第一个基于 ViT 的方法，该方法通过注意力机制增强双模态特征一致性并用边缘信息锐化显著性目标的轮廓。Chen 等人[36]认为 RGB 图像倾向于提供外观和纹理信息，而热力图像以提供几何和空间结构线索为主。受风格迁移的启发，他们提出了一个风格迁移融合模块，在每个层级转换两种图像特征来减少模态的差异性。他们针对融合时高低层特征的平等性问题和空间错位问题设计了一个多尺度通道注意力模块，从全局和局部的角度为多级特征赋权。此外，他们还基于反向注意力来聚合前背景信息，进一步细化目标边界。Tang 等人[37]将研究的视角转向不同于前两者的高分辨率网络，提出了 HRTransNet。基于在网络全程保持高分辨率特征的 HRFormer[38]，他们将热力信息视作辅助模态，通过注意力机制有权重地将其注入到 RGB 模态中，在输入层面上实现特征的融合。此外，他们还利用自注意力和交叉注意力在输出层面上探索了多级特征的互补性。

相较于 CNN 的方法，此类方法在检测完整性上更具优势，这得益于 ViT 捕获远程依赖的能力。但 ViT 将图像分块的操作会导致细节信息丢失，产生块效应（Block Artifact）乃至出现大面积误检区域。事实上，全局信息和局部信息对计算机视觉任务都非常重要，尽管会引入额外的计算，但是利用 CNN 的结构获取局部信息的能力弥补 ViT 的缺陷是一个值得研究的方向。

2 数据集和评价指标

2.1 RGB-T 显著性目标检测数据集

受到设备限制，热红外数据获取困难，早期的

研究提出了一些小规模数据集[8]，但没有被普及使用。Wang 等人提出了第一个 RGB-T 显著性目标检测的基准 VT821[7]。自此以后，RGB-T 显著性目标检测的方法使用的数据集主要以 VT821 为主，部分方法还会加入单独的热红外图和 RGB 图像来扩大训练数据[19]。Tu 带领他的团队，先后又提出了 VT1000[15]和 VT5000[21]，RGB-T 显著性目标检测的数据集得到了极大的扩充，VT821、VT1000 和 VT5000 也成为目前最常用的 RGB-T 显著性目标检测任务的数据集。表 1 简要汇总了这 3 个数据集的相关信息。

表 1 RGB-T 显著性目标检测数据集

Table 1 The RGB-T salient object detection datasets

Name	Year	Scales	Camera equipment	Disadvantage
VT821	2018	821	FLIR A310、SONY TD-2073	1. Simple scenes that lack complexity and variety. 2. The camera uses different parameters when capturing RGB and thermal images. 3. Additional whitespace is introduced when aligning images.
VT1000	2019	1000	FLIR SC620	1. There are potential errors as the images are aligned manually. 2. Limited scenario complexity and diversity.
VT5000	2020	5000	FLIR T640、FLIR T610	1. Images are affected by thermal crossover, making detection challenging.

2.2 常用的评价指标

遵循大多数方法，RGB-T 显著性目标检测使用 PR 曲线、F 测度和平均绝对值误差（Mean Absolute Error，MAE）作为常用的评价指标。此外，RGB-T 显著性目标检测引入了 S 测度（S-measure）和 E 测度（E-measure）来进一步提升评价指标的可靠性。

1）S 测度（S-measure）[39]。S 测度能够评估预测图和真值图之间的结构相似性，并且同时能够评估结构的完整性，S 测度定义为：

$$S_\lambda = \lambda \times S_{\mathrm{O}} + (1-\lambda) S_{\mathrm{r}} \tag{1}$$

式中：S_{r} 为区域感知度量，用于评估局部结构相似性；S_{O} 为对象感知度量，用于解释全局结构相似性，λ 为超参数。

2）E 测度（E-measure）[40]。E 测度是一种模拟人眼判别图片目标的指标，能够捕获自适应的全局和局部相似性，其定义如下：

$$E_{\mathrm{FM}} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi_{\mathrm{FM}}(x, y) \tag{2}$$

式中：h 和 w 分别是图片的高度和宽度；FM 在原文中定义为前景图（foreground map），即检测到的目标。

3 实验分析

本章对近年来部分方法进行了定性和定量比较，旨在更直观地展示 RGB-T 显著性目标检测发展水平。本章采用了 2.2 节中介绍的几个常用的评价指标，在 VT821、VT1000 和 VT5000 数据集上对具有代表性的 RGB-T 显著性目标检测方法进行了定量和定性比较。

3.1 基于机器学习的 RGB-T 显著性检测方法对比

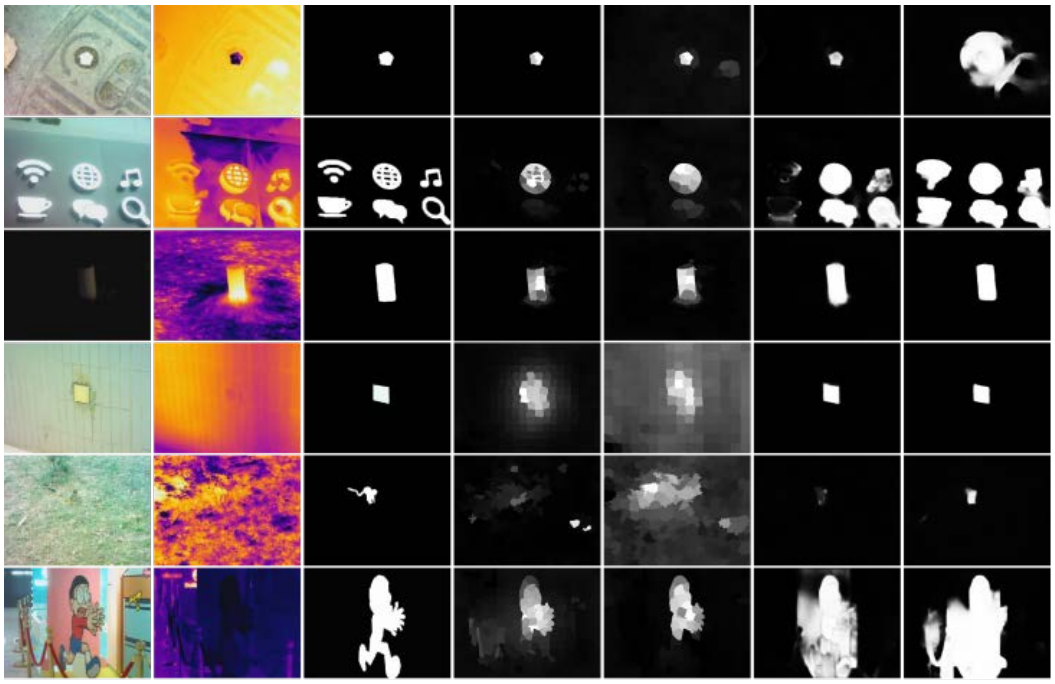
基于机器学习的方法的定量比较结果如表 2 所示。通过表中数据可知，尽管此类方法在几个评价指标上的结果有所提升，但仍然处于一个较低的水平，其可视化结果如图 4（第 4，5 列）所示。从图中可以看出，基于机器学习的方法能够有效地定位出显著性目标区域，但存在大面积的误检、漏检等问题。正如前文所述，由于手工特征的局限性，这类方法只能在一些简单场景产生好的结果（第 1 行），面对复杂情况效果较差甚至失效（第 2～6 行）。由此可见，特征的获取对于 RGB-T 显著性目标检测任务至关重要。

表 2 基于机器学习的 RGB-T 显著性目标检测方法定量比较

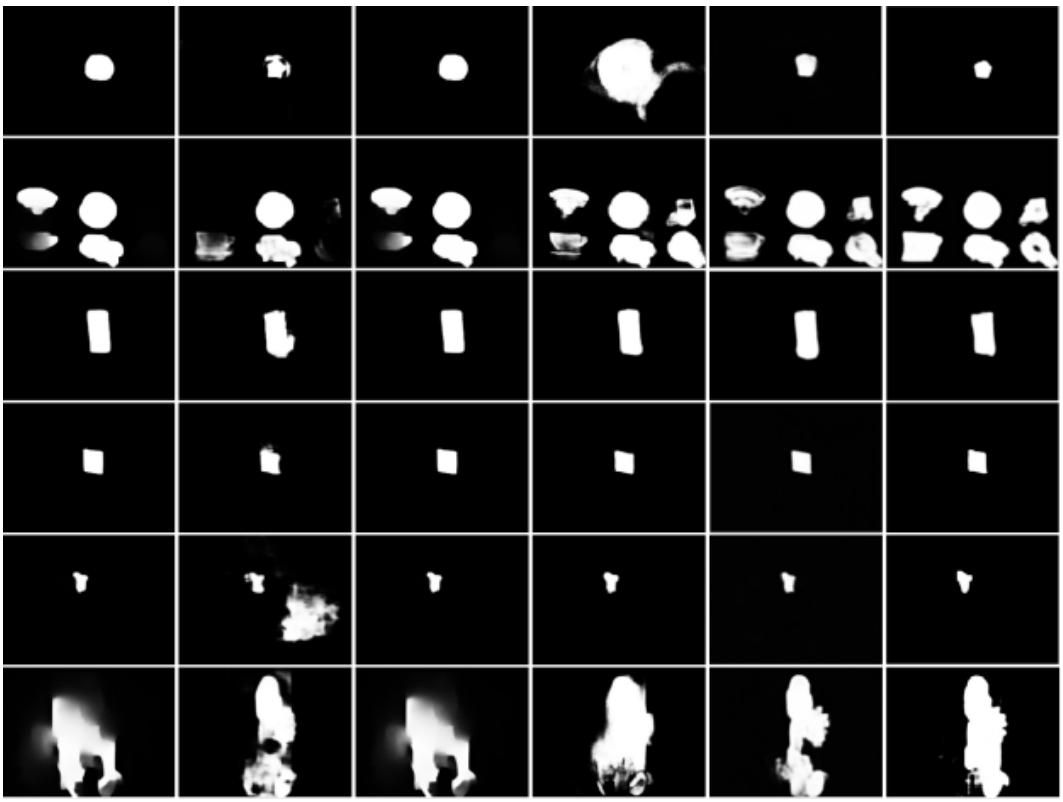
Table 2 Quantitative comparison of machine learning-based RGB-T salient object detection methods

Algorithms	VT821				VT1000				VT5000			
	S↑	F↑	E↑	MAE↓	S↑	F↑	E↑	MAE↓	S↑	F↑	E↑	MAE↓
MTMR[7]	0.725	0.662	0.815	0.108	0.706	0.715	0.836	0.119	0.680	0.595	0.795	0.114
N3S-NIR[10]	0.723	0.734	**0.859**	0.140	0.726	0.717	0.827	0.145	0.652	0.575	0.780	0.168
LTCR[11]	0.762	**0.737**	0.854	0.088	0.799	0.794	0.872	0.084	-			
MGFL[12]	**0.782**	0.725	0.841	**0.071**	**0.820**	**0.801**	**0.882**	**0.066**	**0.751**	**0.661**	**0.817**	**0.085**

Note: ↑ indicates that the larger the indicator, the better, and ↓ indicates that the smaller the indicator, the better. Bold and underline indicate optimal and sub-optimal results, respectively.



(a)RGB (b)Thermal (c)GT (d)MTMR (e)M3S-NIR (f)ADFNet (g)MIDD



(h)CGFNet (i)MMNet (j)CSRNet (k)APNet (l)SwinNet (m)HRTransNet

图 4 RGB-T 显著性目标检测方法的可视化比较

Fig.4 Visual comparison of RGB-T salient object detection methods

3.2 基于深度学习的方法比较

本文将基于深度学习的 RGB-T 显著性目标检测方法分为基于 CNN 的方法和基于 ViT 的方法两类进行实验对比，表 3 展示了基于 CNN 的方法和基于 ViT 的方法的定量比较。由表中数据可知，基于 ViT 的方法在几个数据集的评价指标上大部分都优于基于 CNN 的方法，这可能得益于 ViT 通过建立长期依赖捕获的全局特征相较于 CNN 获取的局部特征包含更

多的有效信息。图 4（第 6～13 列）可视化了部分具有代表性的基于深度学习的方法，包括小目标（第 1 行）、多目标（第 2 行）、低光场景（第 3 行）、低质量热力图像（第 4 行）、难以区分的前景（第 5 行）和杂乱场景（第 6 行）。相比于基于机器学习的方法，基于深度学习的 RGB-T 显著性目标检测方法显著减少了误检和漏检区域。如图 4 所示，在面对多目标时，大部分基于 CNN 的方法存在漏检的问题，而基于 ViT 的方法都能有效地定位出所有的显著目标，这也侧面印证了 ViT 更擅长捕获图像的全局上下文信息。此外，如第 3、4 行所示，低质量的单模态输入对于简单场景下的基于深度学习的 RGB-T 显著性目标检测方法影响较小，但是面对更加复杂的情况（第 5 行），不加以区分地使用双模态特征会显著地影响预测结果（第 6、9 列）。此外，尽管 RGB 图像和热力图像都提供了充分的细节，但根据第 2 行的结果，所有的方法都没有很好地刻画出每个目标镂空部分的轮廓。第 5 行所展示场景中，大部分方法也同样没有能够定位出完整目标边界。事实上，边界模糊的问题是包括 RGB-T 显著性目标检测在内的所有像素级预测任务都面临的挑战。随着成像技术的发展，现在获取到的图像分辨率极高，例如清晰度为 1080p 的图像大约包含 200 万个像素。目前的视觉任务为了减少网络的计算量，通常会将输入图像调整为合适的大小（如 224×224，256×256 等），在网络处理完之后通过上采样的方式恢复分辨率，但这种操作会不可逆地丢失一部分信息。为此，现有的方法通常会在解码阶段引入浅层的编码特征来补充细节信息，但这不可避免地引入了噪声。此外，CNN 为了减少计算和扩大感受野而交替使用卷积层和池化层，这进一步加剧了不可逆的信息丢失。

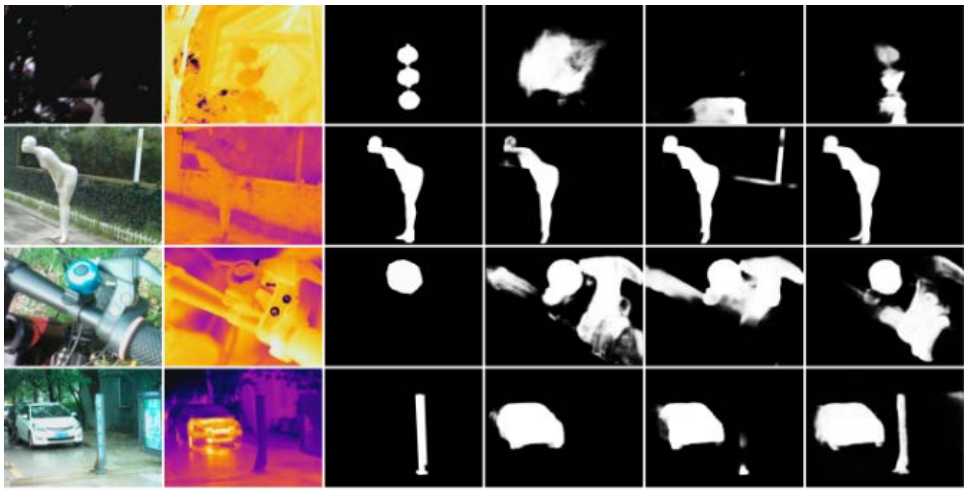
除上述问题外，目前 RGB-T 显著性目标检测还面临着以下挑战：①低质量的 RGB 图像和低质量的热红外图。如图 5 所示，在低光环境下（第 1 行）RGB 图像无法提供足够的颜色、纹理等信息，而低质量的热力图又存在大量干扰信息，现有的方法无法准确地定位显著性目标。②数据集误差。现有的 RGB-T 显著性目标检测数据集中存在着未对齐的掩膜（第 2 行）和有歧义的注释（第 3～4 行）等问题，这会误导模型使其无法得到正确的预测结果。

表 3 基于深度学习的 RGB-T 显著性目标检测方法定量比较

Table 3 Quantitative comparison of deep learning-based RGB-T salient object detection methods

Methods	Algorithms	Backbone	VT821				VT1000				VT5000			
			S↑	F↑	E↑	MAE↓	S↑	F↑	E↑	MAE↓	S↑	F↑	E↑	MAE↓
CNN-based	FMCF[8]	VGG16	0.760	0.640	0.796	0.080	0.873	0.823	0.921	0.037	0.814	0.734	0.864	0.055
	SGDL[15]	VGG19	0.765	0.730	0.847	0.085	0.787	0.764	0.856	0.090	0.750	0.672	0.824	0.089
	ADFNet[21]	VGG16	0.810	0.716	0.842	0.077	0.910	0.847	0.921	0.034	0.863	0.778	0.891	0.048
	MIDD[22]	VGG16	0.871	0.804	0.895	0.045	0.915	0.882	0.933	0.027	0.867	0.801	0.897	0.043
	CGFNet[23]	VGG16	0.881	0.845	0.912	0.038	0.923	**0.906**	0.944	0.023	0.883	0.851	0.922	0.035
	CGMDRNet[25]	Res2Net-50	0.894	0.840	0.920	0.035	0.931	0.893	0.940	0.020	0.896	0.846	0.928	0.032
	TNet[27]	ResNet-50	0.898	0.841	0.919	0.030	0.928	0.889	0.937	0.021	0.894	0.847	0.927	0.033
	MIA_DPD[28]	ResNet-50	0.844	-	0.850	0.070	0.924	-	0.926	0.025	0.879	-	0.893	0.040
	MMNet[29]	ResNet-50	0.875	0.798	0.893	0.040	0.917	0.863	0.924	0.027	0.864	0.785	0.890	0.043
	CAVER[30]	ResNet-50	0.891	0.839	0.919	0.033	0.935	0.903	0.945	0.018	0.891	0.842	0.930	0.032
	CSRNet[31]	ESPNet’v2	0.885	0.830	0.908	0.038	0.918	0.877	0.925	0.024	0.868	0.810	0.905	0.042
ViT-based	SwinNet[35]	Swin transformer	0.904	0.847	0.926	0.030	**0.938**	0.896	0.947	0.018	**0.912**	0.865	0.942	0.026
	HRTransNet[37]	HRFormer	**0.906**	**0.853**	**0.929**	**0.026**	**0.938**	0.900	0.945	0.017	**0.912**	**0.871**	**0.945**	**0.025**
	MITF-Net[36]	PVT’v2	0.905	**0.853**	0.927	0.027	**0.938**	**0.906**	**0.949**	**0.016**	0.910	0.870	0.943	**0.025**

Note: ↑ indicates that the larger the indicator, the better, and ↓ indicates that the smaller the indicator, the better. Bold and underline indicate optimal and sub-optimal results, respectively.



(a)RGB (b)Thermal (c)GT (d)APNet (e)CGFNet (f) SwinNet

图 5 RGB-T 显著性目标检测面临的挑战

Fig.5 The challenges faced by RGB-T salient object detection

4 结语

本文整理了现有的RGB-T显著性目标检测方法，分为机器学习的方法和基于深度学习的方法两大类进行总结，并从结构的角度着重分析了基于深度学习的方法。此外，本文还归纳了 RGB-T 显著性目标检测常用的数据集和具有代表性的评价指标，并在其之上对现有的方法进行实验分析。与机器学习的方法相比，基于深度学习的方法能够从大量的数据中提取不同的富含丰富信息的特征，具有更强的泛化能力，更善于应对不同的复杂场景。但该领域仍然面临着一些诸如边界模糊、低质量输入等方面的挑战，对于双模态特征互补性、边界信息的探索仍然值得继续深挖。对于 RGB-T 显著性目标检测任务的未来发展，可从以下两方面考虑：

1）数据集：如前文所述，现有的RGB-T显著性目标检测常用的数据集仍然存在一些问题。这可能和数据采集方式、标注显著性目标的主观性等方面有关。此外，诸如显著性目标检测、显著性实例分割等相近领域拥有数量更多、规模更大、场景更加多样的数据集[41-42]，能够为训练模型的泛化能力提供数据上的支持。目前 RGB-T 显著性目标检测 3 个常用的数据集中，有且仅有一个规模相对较大的数据集VT5000。而数据的数量和质量能够明显地影响模型的好坏，更加准确有效的大规模数据集对 RGB-T 显著性目标检测具有重要的研究意义。

2）监督方式：目前的RGB-T显著性目标检测模型大多基于像素级全监督信息，但在实际场景中，大规模、高质量的像素级标注的成本是巨大的。因此，探究在少量甚至无标注数据下的 RGB-T 显著性目标检测方法是非常有价值的，比如弱监督/半监督/自监督等方法的研究。近期，Liu 等人[14]基于涂鸦提出了第一个弱监督的 RGB-T 显著性目标检测方法，为该方向接下来的发展提供了借鉴。

参考文献：

[1] XU H, ZHANG H, MA J Y. Classification saliency-based rule for visible and infrared image fusion[J]. *IEEE Transactions on Computational Imaging*, 2021, **7**: 824-836.

[2] LI G Y, WANG Y K, LIU Z, et al. RGB-T semantic segmentation with location, activation, and sharpening [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(3): 1223-1235.

[3] 侯毅苇, 李林汉, 王彦. 结合红外显著性目标导引的改进 YOLO 网络的智能装备目标识别研究[J]. **红外技术**, 2020, **42**(7): 644-650.
HOU Yiwei, LI Linhan, WANG Yan. Intelligent equipment object recognition based on improved YOLO network guided by infrared saliency detection[J]. *Infrared Technology*, 2020, **42**(7): 644-650.

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

[5] LI C L, CHENG H, HU S Y, et al. Learning collaborative sparse representation for grayscale-thermal tracking[J]. *IEEE Transactions on Image Processing*, 2016, **25**(12): 5743-5756.

[6] 张骏, 张鹏, 张政, 等. 类 HED 网络的热红外图像显著性人体检测深度网络[J]. **红外技术**, 2023, **45**(6): 649-657.
ZHANG Jun, ZHANG Peng, ZHANG Zheng, et al. Similar HED-Net for salient human detection in thermal infrared images[J]. *Infrared Technology*, 2023, **45**(6): 649-657.

[7] WANG G Z, LI C L, MA Y P, et al. RGB-T saliency detection benchmark: dataset, baselines, analysis and a novel approach[C]//IGTA 2018: *The* 13th *Academic Conference on Image Graphics Technology and Application*, 2018: 359-369.

[8] MA Y, SUN D, MENG Q, et al. Learning multiscale deep features and svm regressors for adaptive RGB-T saliency detection[C]//ISCID 2017: 2017

10th *International Symposium on Computational Intelligence and Design*, 2017: 389-392.

[9] ZHOU D Y, Weston J, Gretton A, et al. Ranking on data manifolds[C]// NIPS 2003: *Advances in Neural Information Processing Systems*, 2003: 169-176.

[10] TU Z Z, XIA T, LI C L, et al. M3S-NIR: multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection[C]// MIPR 2019: 2019 *IEEE Conference on Multimedia Information Processing and Retrieval*, 2019: 141-146.

[11] HUANG L M, SONG K C, WANG J, et al. Multi-graph fusion and learning for RGBT image saliency detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(3): 1366-1377.

[12] HUANG L M, SONG K C, GONG A J, et al. RGB-T saliency detection via low-rank tensor learning and unified collaborative ranking[J]. *IEEE Signal Processing Letters*, 2020, **27**: 1585-1589.

[13] 张冬明, 靳国庆, 代锋, 等. 基于深度融合的显著性目标检测算法[J]. **计算机学报**, 2019, **42**(9): 2076-2086.
ZHANG D M, JIN G Q, DAI F. Sailent object detection based on deep fusion of hand-craft features[J]. *Chinese Journal of Computers*, 2019, **42**(9): 2076-2086.

[14] Sandler M, Howard A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// CVPR 2018: *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 4510-4520.

[15] TU Z Z, XIA T, LI C L, et al. RGB-t image saliency detection via collaborative graph learning[J]. *IEEE Transactions on Multimedia*, 2020, **22**(1): 160-173.

[16] PANG Y, WU H, WU C D. Cross-modal co-feedback cellular automata for RGB-T saliency detection[J]. *Pattern Recognition*, 2023, **135**: 109-138.

[17] LIU Z Y, HUANG X S, ZHANG G H et al. Scribble-supervised RGB-T salient object detection[C]//ICME 2023: *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2023: 2369-2374.

[18] ZHANG Q, HUANG N C, YAO L, et al. RGB-T salient object detection via fusing multi-level CNN features[J]. *IEEE Transactions on Image Processing*, 2020, **29**: 3321-3335.

[19] ZHANG Q, HUANG N C, XIAO T, et al. Revisiting feature fusion for RGB-T salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, **31**(5): 1804-1818.

[20] BI H B, WU R W, LIU Z Q, et al. PSNet: parallel symmetric network for RGB-T salient object detection[J]. *Neurocomputing*, 2022, **511**: 410-425.

[21] TU Z Z, MA Y, LI Z, et al. RGBT salient object detection: a large-scale dataset and benchmark[J]. *IEEE Transactions on Multimedia*, 2022, **25**: 4163-4176.

[22] TU Z Z, LI Z, LI C L, et al. Multi-interactive dual-decoder for RGB-thermal salient object detection[J]. *IEEE Transactions on Image Processing*, 2021, **30**: 5678-5691.

[23] WANG J, SONG K C, 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*, 2022, **32**(5): 2949-2961.

[24] CHEN Q, LIU Z, ZHANG Y, et al. RGB-D Salient Object Detection via 3D Convolutional Neural Networks[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022: 1063-1071.

[25] CHEN G, SHAO F, CHAI X L, et al. CGMDRNet: cross-guided modality difference reduction network for RGB-T salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(9): 6308-6323.

[26] LIAO G B, GAO W, LI G, et al. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(11): 7646-7661.

[27] CONG R M, ZHANG K P, ZHANG C, et al. Does thermal really always matter for RGB-T salient object detection?[J]. *IEEE Transactions on Multimedia*, 2022, **25**: 1-12.

[28] LIANG Y H, QIN G H, SUN M H, et al. Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection[J]. *Neurocomputing*, 2022, **490**: 132-145.

[29] GAO W, LIAO G B, MA S W, et al. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(4): 2091-2106.

[30] PANG Y W, ZHAO X Q, ZHANG L H, et al. CAVER: cross-modal view-mixed transformer for bi-modal salient object detection[J]. *IEEE Transactions on Image Processing*, 2023, **32**: 892-904.

[31] ZHOU W J, GUO Q L, LEI J S, et al. ECFFNet: effective and consistent feature fusion network for RGB-T salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(3): 1224-1235.

[32] ZHOU W J, ZHU Y, LEI J S, et al. LSNet: lightweight spatial boosting network for detecting salient objects in RGB-thermal images[J]. *IEEE Transactions on Image Processing*, 2023, **32**: 1329-1340.

[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//NIPS 2017: *Advances in Neural Information Processing Systems*, 2017: 6000-6010.

[34] WANG W H, XIE E Z, LI X, et al. PVTv2: Improved baselines with pyramid vision transformer[J]. *Computational Visual Media*, 2021, **8**: 415-424.

[35] LIU Z Y, TAN Y C, 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*, 2022, **32**(7): 4486-4497.

[36] 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.

[37] TANG B, LIU Z Y, TAN Y C, et al. HRTransNet: HRFormer-driven two-modality salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(2): 728-742.

[38] YUAN Y H, FU R, HUANG L, et al. HRFormer: high-resolution vision transformer for dense predict[C]//NIPS 2021: *Advances in Neural Information Processing Systems*, *Virtual*, 2021: 7281-7293.

[39] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: a new way to evaluate foreground maps[C]//ICCV 2017: *Proceedings of the* 2017 *IEEE/CVF International Conference on Computer Vision*, 2017: 4558-4567.

[40] FAN D P, GONG C, CAO Y, et al. Enhanced-alignment measure for binary foreground map evaluation[C]//IJCAI 2018: The 27th *International Joint Conference on Artificial Intelligence*, 2018: 698-704.

[41] YAN Q, XU L, SHI J P, et al. Hierarchical saliency detection[C]//CVPR 2013: *Proceedings of the* 2013 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2013: 1155-1162.

[42] LIN Y, HOU X D, Koch C, et al. The secrets of salient object segmentation[C]//CVPR 2014: *Proceedings of the* 2014 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014: 280-287.