基于双边注意力机制的 RGB-D 显著性物体检测方法 (中译版)

张钊, 林铮, 徐君, 金闻达, 卢少平, 范登平

摘要

RGB-D 显著性物体检测（SOD）致力于在一对跨膜态 RGB 和深度图像中分割出最吸引人的物体。目前，在利用深度图像时，大多数现有的 RGB-D SOD 方法都集中在前景区域。然而，背景区域也为传统的 SOD 方法提供了重要的信息，以实现更好的检测性能。

为了更好地探索前景和背景区域中的显著信息，本文针对 RGB-D SOD 任务提出了一个双边注意网络（BiANet）。具体来说，我们引入了一种具有互补注意力机制的双边注意力模块（BAM:Bilateral Attention Module）：前景优先（FF:foreground-first）注意和背景优先（BF:background-first）注意。

前景优先（FF）的注意力逐渐集中在前景区域，而背景优先（BF）在背景区域中恢复可能有用的显著信息。

受益于引入的 BAM 模块，我们的 BiANet 可以捕获更有意义的前景和背景线索，并将更多的注意力转移到细化前景和背景区域之间的不确定细节上。

此外，我们利用多尺度技术扩展了 BAM 模块，以获得更好的 SOD 性能。

在六个基准数据集上进行的广泛实验表明，在客观指标和主观视觉比较方面，我们的 BiANet 性能优于其他最新的 RGB-D SOD 方法。

使用英伟达 GeForce RTX 2080Ti GPU，我们提出的 BiANet 在 224×224 RGB-D 图片上运行最高可以达到 80fps

全面的消融实验也验证了我们的贡献。

关键词—双边注意力, **显著性物体检测**, RGB-D **图片**.

BiANet 相关代码开源在https://github.com/zzhanghub/bianet 张钊, 林铮, 徐君, 卢少平和范登平属于 TKLNDST, 在天津南开大学计算机学院。金闻达在天津大学智能与计算学部。卢少平为相关作者, （邮箱: slu@nankai.edu.cn）。本研究得到新一代人工智能重大专项的支持, 基金号: 2018AAA0100400, 国家自然科学基金 (61922046、61972216), 国家青年人才支持计划, 和天津市自然科学基金 (17JCJQJC43700, 18JCYBJC41300)。

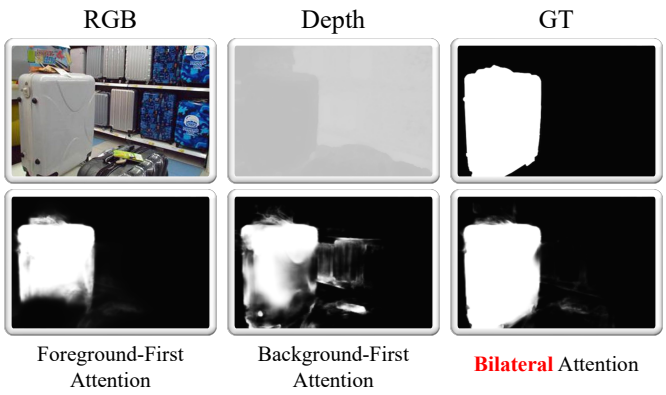

图 1. **前景优先、背景优先和我们的双边关注机制的** RGB-D SOD **结果对比**。深度信息提供了丰富的前景和背景关系。更加关注前景有助于预测高可信度的前景对象，但可能会产生不完整的结果。更加关注背景可以找到更完整的对象，但是可能会引入其他噪音。我们的 BiANet 联合探索前景和背景线索，在背景噪声很小的情况下实现了完整的前景预测。

I. 简介

为了实时理解复杂场景，人类能够在进一步处理之前从所有可用视觉信息种过滤出视觉上独特的部分，也就是所谓的显著性物体 [33], [69]。生理学，认知心理学，计算机视觉等领域内专家已经就人类这项能力进行了长期研究 [10], [32], [90]。一个显著的物体可以在颜色，形状，距离上与周围物体区分开 [4], [47]。事实证明，在广泛的视觉应用种，首先捕获吸引注意力的对象是有效的。，例如视觉追踪 [41], [49]，图片分割 [31], [36], [68], 视频分析和检测 [18], [74], [84], 图像检索 [46], 图像协同分割 [15], [16], [86]。现存的大多数 SOD 方法 [34], [48], [82] 主要应用于 RGB 图像。然而，它们在纹理相似，背景复杂或对象均质的情况下通常会产生不准确的 SOD 结果 [73], [83].

随着智能手机中深度传感器的普及，与相应 RGB 图像关联的深度图变得越来越容易获取。直观地，深度信息（例如 3D 布局和空间线索）对于减少 RGB 图像中的歧义至关重要，同时也是改善 SOD 性能的重要补充 [38]。因此，RGB-D SOD 受到越来越多研究者的关注 [7], [19], [23], [25], [61], [78], [80], [81]。

对于当前的 RGB-D SOD 方法，深度对比度是最

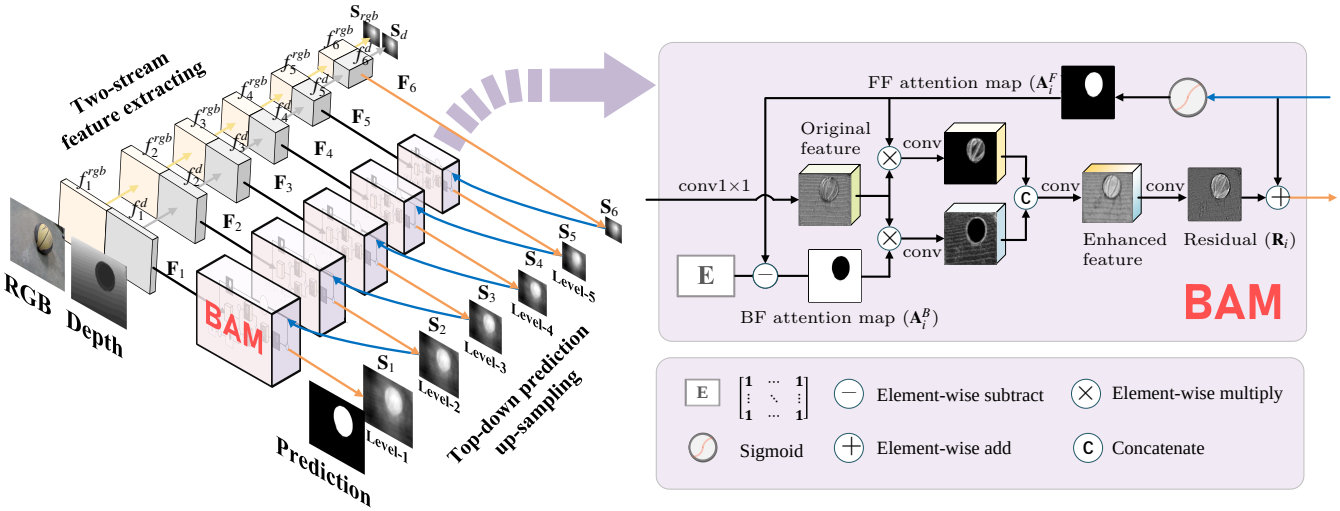

图 2. BiANet **的整体架构**. BAM 即为我们提出的双边注意力模块，同时也可选择性地使用它的多尺度扩展（MBAM）作为替代。BiANet 包含三个主要步骤: 双流（two-stream）特征提取, 自上而下的预测上采样, 和双边注意力残差补偿 (通过 BAM)。具体地，模型首先从图像的 RGB 和深度流中提取多尺度特征 $\{f_i^{rgb}, f_i^d\}_{i=1}^6$，然后，把他们进行拼接得到 $\{\mathbf{F}_i\}_{i=1}^6$。从 f_6^{rgb} and f_6^d 预测得到 $\mathbf{S}_{rgb}$ 和 $\mathbf{S}_d$ 显著图以备深度监督学习。我们使用顶层特征 $\mathbf{F}_6$ 来预测出一个粗糙的显著图 $\mathbf{S}_6$。为了获得精准、高分辨的结果，我们对初始特征图进行上采样，同时采取自上而下的方式使用 BAM 补偿显著图中的细节。BAMs 同时接受高维预测结果 $\mathbf{S}_{i+1}$ 和当前维度特征 $\mathbf{F}_i$ 作为输入。在一个 BAM 中，可以根据 $\mathbf{S}_{i+1}$ 计算出前景优先注意力图 $\mathbf{A}_i^F$ 和背景优先注意力图 $\mathbf{A}_i^B$。我们应用双重互补注意图来探索前景和背景线索，共同推断出残差项以细化上采样过的显著图。We apply the dual complementary attention maps to explore the foreground and background cues bilaterally, and jointly infer the residual for refining the up-sampled saliency map.

重要的先验项，它通常用于在与背景区域有强烈对比的前景区域上转移更多优先级。For current RGB-D SOD methods, the depth contrast has served as the most important prior [59], [64], [66], [87], and it is often used to shift more priority on the foreground regions which have a strong contrast with the background. 例如，在早期的 RGB-D SOD 工作中，Fan 等人 [20] 使用深度图作为颜色对比度的权重因子。CPFP 的最新工作 [87] 设计了一个有效性损失来增强深度对比度，从而更好地使网络专注于前景区域。 对前景区域的更多关注确实有助于学习到显著性线索。同时，像这些工作 [44], [75], [76] 中所展示的，了解场景中的背景信息也可以帮助提高 SOD 性能。前景先验和背景先验有很大不同。例如，前景先验包含更多吸引人类视觉注意力的线索，例如敏感类别（sensitive categories），明亮的颜色，特殊形状，与观察者的距离更近，而背景（非显著）先验则相反。因此，有必要分别探索前景和背景线索，然后共同在场景中挖掘更准确的显著区域。

一些传统方法也会使用这种方式预测显著性物体。受益于共同探索前景和背景线索，这些方法在当时达到了领先的效果。然而，当前的 RGB-D SOD 网络很大程度上忽略了这种简单有效的想法。在本文中，我们提出了一个双边注意力网络（BiANet），以便从 RGB 流和深度流中共同学习互补的前景和背景特征，以获得更好的 RGB-D SOD 性能。如图图2所示，我们的 BiANet 采用两流架构，同时 RGB 流和深度流的输出在多阶段都被接在一起。and the side outputs from the RGB and depth streams are concatenated in multiple stages. 首先，我们使用高级语义特征 $\mathbf{F}_6$ 来定位前景和背景区域 $\mathbf{S}_6$。然而，初始的显著图 $\mathbf{S}_6$ 是粗糙、低分辨率的。为了增强粗糙显着性图，我们设计了一个双边注意力模块（BAM），该模块由互补的前景优先（FF）注意和背景优先（BF）注意机制组成。FF 将注意力转移到前景区域上，以逐渐完善显著性预测，而 BF 则专注于背景区域，以恢复边界附近的潜在显著区域。如图所示图1，通过双向探索前景和背景线索，该模型有助于更准确地进行预测。其次，我们提出了 BAM 的多尺度扩展（MBAM），可以有效地学习多尺度的上下文信息，捕获局部和全局的显著性信息，以进一步提高 SOD 性能。在六个基准数据集上进行的大量实验表明，我们的 BiANet 比以前的 RGB-D SOD 技术具有更好的性能，同时简单的体系结构使我们的模型非常快。

总而言之，我们的贡献主要有三方面:

- **我们提出了一个简单而有效的双边注意力模块**（BAM），以便与深度图像中丰富的前景和背景信息共同探索前景和背景线索。与最新方法相比，我们的 BiANet 在 9 个标准指标下在 6 个流行的 RGB-D SOD 数据集上实现了更好的性能，同时拥

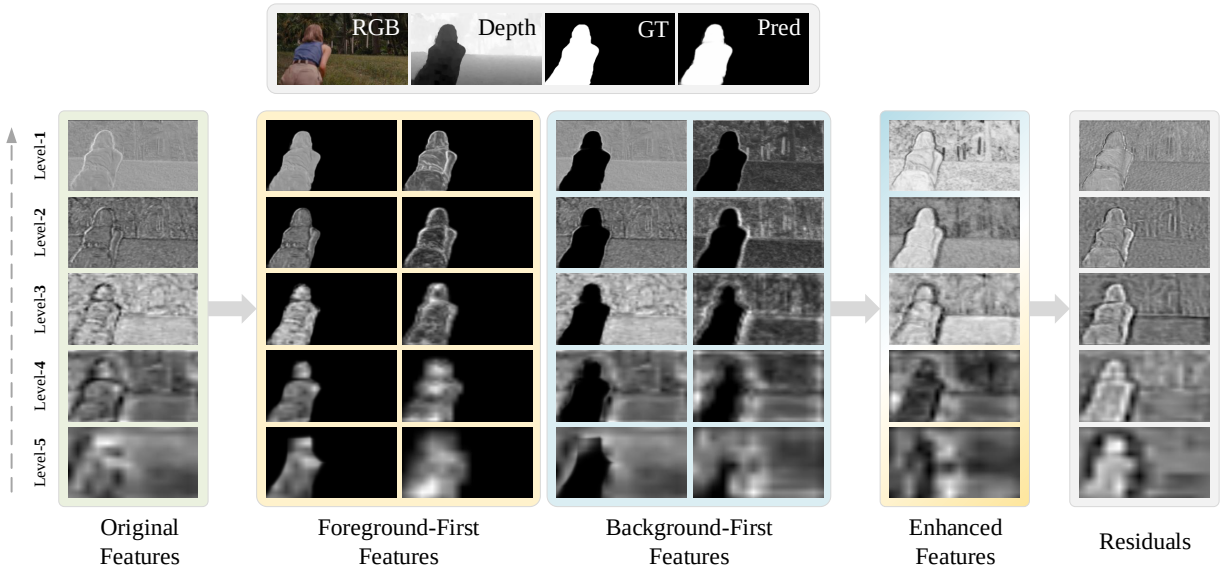

图 3. **双边注意力工作机制的可视化**. 原始特征是每个层级中的平均边缘输出的特征。The original features are the averaged side-output features in each level. 我们在黄色和蓝色框的左列中显示出原始特征直接乘以前景/背景优先注意力图的结果。这两个方框的右列是在两个分支中进一步卷积后的特征。可以看出，前景优先特征关注前景区域以探索显著性线索；而背景优先特征将更多注意力转移到背景区域以挖掘潜在的重要对象。无论是前景优先特征还是背景优先特征，都将更多优先级转移到由上采样引起的不确定区域(low confidence)。当融合两个分支并共同进行推断时，我们可以看到双边增强特征对前景或背景在哪里有了更准确的理解。由于获得更多关注，不确定区域可以被具有强烈对比度的残差项重新分配给正确的属性。Due to obtaining more attention, the uncertain areas are reassigned to the right attribution by the residual with strong contrast. "Pred" 是模型的预测结果。'Pred' is the prediction of the model.

有更好的视觉效果（例如，包含更多的细节和锐利的边缘）。

- **与最新方法相比，我们的** BiANet **在** 9 **个标准指标下在** 6 **个流行的** RGB-D SOD **数据集上实现了更好的性能**，同时拥有更好的视觉效果（例如，包含更多的细节和锐利的边缘）。在不同设置下，我们的 BiANet 都可以在 NVIDIA GeForce RTX2080Ti GPU 上以 34 fps ~80fps 运行，对于实际应用而言，这是一个可行的解决方案。
- **在不同设置下**，**我们的** BiANet **都可以在** NVIDIA GeForce RTX2080Ti GPU **上以** 34 fps ~80fps **运行**，对于实际应用而言，这是一个可行的解决方案。

本文的其余部分结构如下。

在 §II 部分，我们简要梳理相关工作。

在 §III部分，我们提出用于 RGB-D 显著性物体检测的双边注意力网络（BiANet）。

在 §IV 部分，我们在六个基准数据集上进行了广泛的实验，并与当前最新的 RGB-D SOD 方法相比较来评估模型的性能。

总结部分在 §V。The conclusion is given in §V.

II. 相关工作

A. RGB-D 显著性物体检测

RGB-D 显著性物体检测（SOD）致力于在一对跨膜态 RGB 和深度图像中分割出最吸引人的物体。早期方法主要集中于从 RGB 和深度图像中提取低级显著性线索，探索物体距离 [38], 高斯差异 [35], 图信息 [12], 多层次判别显著融合 [66], 多上下文对比 [11], [59], 背景闭合（background enclosure） [21], etc. 但是，由于缺少高级特征表示，这些方法很容易导致显着性预测失效。最近，Qu 等人 [63] 引入了深度神经网络（DNN）来研究多种显著性线索的高级表示形式，包括局部和全局对比度以及色彩聚集度（color compactness）。之后，使用 DNN 查找 RGB 和深度图像的高级表示形式已经被广泛 [7], [24], [39], [79] 应用在在 RGB-D SOD 任务中。例如，一些工作 [8], [28], [70] 尝试分两步先分别提取 RGB 和深度特征，然后在网络的浅层、中层或深层中将两者合并。通过替代一次性集成为多级阶段融合交叉模式特征，现有方法 [6], [7], [40], [61] 进一步提高了 SOD 性能。Fan 等人 提出深度图并不总是有利于显着物体检测 [17]；因此，他们提出了一种深度净化单元来自动丢弃一些劣质的深度图。

B. 前景线索和背景线索

前景和背景的分布差异很大，因此有必要探讨它们各自的线索。在传统方法中，有些工作集中于共同推断前景和背景中的显著区域。Yang 等人 提出了一种两阶段的 SOD 方法 [77] 。它首先将输入图像的顶部，底部，左侧和右侧边缘区域视为背景种子，通过基于图形的流形排序来推断可能的前景超像素。然后，根据前景种子对图形进行最终预测排名。Ren 等人 [44] 采用边界连通性来定位初始背景区域，而不是仅假设边界为背景。Liang 等人 [44] 引入了深度图来将远离观察者的区域作为初始背景区域。

III. RGB-D SOD 任务新模型 BiANet

在本节中, 我们首先介绍 BiANet 的总体架构, 然后介绍双边注意模块（BAM）及其多尺度扩展（MBAM）。

A. 结构总览

如 图2中所示，我们的双边注意力网络（BiANet）包含三个主要步骤：特征提取，预测上采样和双边注意力残差补偿。我们从 RGB 和深度流中提取多级特征。随着网络深度的增加，高级特征 (e.g., $\mathbf{F}_4$) 将更有效地捕获全局上下文，但会丢失对象的细节信息。当我们对高级预测进行上采样时，显著图将变得模糊 (e.g., $\mathbf{S}_5$)，并且边缘会变得很难找到。也就是说，在 Sigmoid 层之后，像素位置的预测值接近 0.5。因此，我们使用提出的的双边注意力模块（BAM）将这些不确定区域区分为前景或背景。

1) 特征提取: 我们用两个流对 RGB 和深度信息进行编码。具体来说，RGB 流和深度流均采用 VGG-16 [65] 中的五个卷积模块作为标准 backbone 并附加一个带有三个卷积层的卷积组，来分别预测显著图。不同于之前的工作 [8], [28], [91], 我们在多个阶段探索 RGB 和深度特征的交叉模式融合，而不是在低阶或高阶中将它们只融合一次。 RGB 流的第 i 个边缘输出 f_i^{rgb} 和深度流中的 f_i^d 被结合成一个特征张量 $\mathbf{F}_i$ (tensor)。现在 $\mathbf{F}_6$ 是通过 $M(f_5^{rgb})$ 和 $M(f_5^d)$ 拼接得来, $M(\cdot)$ 代表最大池化操作。 使用两个 3×3 的卷积层通过 $\mathbf{F_6}$ 预测出粗糙的显著图 $\mathbf{S}_6$, 同时 $\{\mathbf{F}_1,\mathbf{F}_2,\cdots,\mathbf{F}_5\}$ 也为 BiANet 中的 BAMs 做好了准备，以通过自上而下的方式将不确定区域区分为前景或背景，从而进一步完善上采样的显著图。

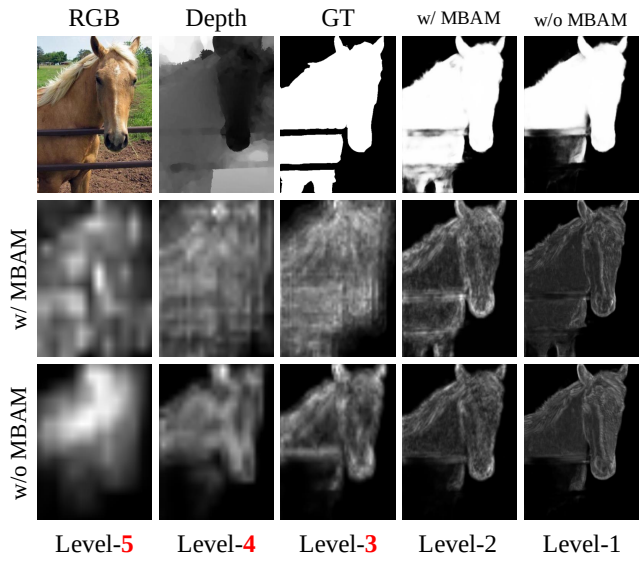

图 4. **由 MBAM 和 BAM 捕获到的高级特征间的区别**。第二行是模型中的平均前景优先特征，其中 MBAM 被应用于前三层（由红色数字标识）。第三行是从模型获得的平均前景优先特征，在该模型中，所有层都配备了 BAM。我们可以看到，与应用 BAM 相比，更高级别的 MBAM 可以捕获更完整的信息，如第一行所示，有利于目标定位。

2) 预测的上采样: 从高级特征预测的初始显著图在低分辨率下较粗糙，但由于它包含丰富的语义信息，因此可用于预测前景和背景的初始位置。为了细化初始显著图 $\mathbf{S}_6$，我们使用具有更多细节信息的较低级别特征 $\mathbf{F}_5$ ，同时在 BAM 的帮助下，预测较高级别的输出与真实标签 (**GT**) 之间的残差部分。我们将预测的残差部分 $\mathbf{R}_5$ 添加到上采样后的更高级别的预测 S_6，获得细化后的预测结果 $\mathbf{S}_5$ ，以此类推，

$$\mathbf{S}_i = \mathbf{R}_i + U(\mathbf{S}_{i+1}), i \in \{1,\ldots,5\}, \tag{1}$$

其中 $U(\cdot)$ 代表上采样。最后，我们的 BiANet 通过 $\mathbf{S} = \sigma(\mathbf{S}_1)$ 获得显著图，其中 $\sigma(\cdot)$ 是 Sigmoid 函数。

3) 双边注意力残差补偿: 为了获得更好的残差并区分上采样后的前景和背景区域，我们设计了一个双边注意力模块（BAM），来让我们的 BiANet 能够区分前景和背景。在我们的 BAM 中，较高级别的预测用作前景优先（FF）注意力图，反向预测作为背景优先（BF）注意力图，将两边对前景和背景的注意力结合起来。在图图3中，可以看到 BAM 生成的残差在对象边界处具有很高的对比度。III-B 和 III-C部分中介绍了更多细节。

4) 损失函数: 深度监督在 SOD 任务中被广泛使用 [22], [30]。它阐明了网络每个步骤的优化目标，并加速了训练的收敛。为了快速收敛，我们还在深度流输出 $\mathbf{S}_d$，RGB 流输出 $\mathbf{S}_{rgb}$ 和每一个由上至下的边缘输出

$\{\mathbf{S}_1,\mathbf{S}_2,\cdots,\mathbf{S}_6\}$ 中应用了深度监督。我们的 BiANet 模型的整体损失函数为:

$$\begin{aligned}\mathcal{L}=&\sum_{i=1}^{6} w_i\mathcal{L}_{ce}\left(\sigma\left(\mathbf{S}_i\right),\mathbf{GT}\right)+w_d\mathcal{L}_{ce}\left(\sigma\left(\mathbf{S}_d\right),\mathbf{GT}\right)\\&+w_{rgb}\mathcal{L}_{ce}\left(\sigma\left(\mathbf{S}_{rgb}\right),\mathbf{GT}\right),\end{aligned}\tag{2}$$

其中 w_i, w_d, and w_{rgb} 代表权重参数，在实验中我们简单设定为 1。$\mathcal{L}_{ce}(\cdot)$ 是二元交叉熵损失，公式为:

$$\mathcal{L}_{ce}(\mathbf{X},\mathbf{Y})=-\frac{1}{N}\sum_{i=1}^{N}\Big(y_i log(x_i)+(1-y_i)log(1-x_i)\Big).\tag{3}$$

上述公式中, $x_i\in\mathbf{X}$ 、$y_i\in\mathbf{Y}$, 和 N 代表全部的像素数。

B. 双边注意力模块 (BAM)

给定最初的前景和背景，如何使用高分辨率的交叉模态特征来细化预测输出是本文的重点。考虑到前景和背景的分布有很大不同，我们使用一对反向注意力组件来设计双边注意力模块，分别从前景和背景中学习特征，然后共同完善预测。如图所示 图2, 为了专注于前景，我们使用被 sigmoid 激活后，从高层上采样后的预测结果作为前景优先（FF）注意力图 $\{\mathbf{A}^F\}_{i=1}^5$，同时背景优先（BF）注意力图 $\{\mathbf{A}^B\}_{i=1}^5$ 是通过从矩阵 $\mathbf{E}$（全部是 1）中减去 FF 图生成的。也就是说,

$$\begin{cases}\mathbf{A}_i^F=\sigma\Big(U(\mathbf{S}_{i+1})\Big),\\ \mathbf{A}_i^B=\mathbf{E}-\sigma\Big(U(\mathbf{S}_{i+1})\Big),\end{cases}\quad i\in\{1,2,3,4,5\}.\tag{4}$$

然后, 由 图2 所示，我们分别应用 FF 和 BF 加权两个分支的边缘输出特征，并进一步联合预测残差分量。

$$\mathbf{R}_i=\mathcal{P}_i^R\left(\left[\mathcal{P}_i^F\left(\hat{\mathbf{F}}_i\odot\mathbf{A}_i^F\right),\mathcal{P}_i^B\left(\hat{\mathbf{F}}_i\odot\mathbf{A}_i^B\right)\right]\right).\tag{5}$$

在上式中, $\odot$ 代表逐元素相乘。$\hat{\mathbf{F}}_i$ 是使用 32 个 1×1 卷积来减少计算开销的通道简化版 $\mathbf{F}_i$。$\mathcal{P}_i^F$ 和 $\mathcal{P}_i^B$ 是前景优先和背景优先分支。他们由 32 个 3×3 的卷积核和一个 ReLU 层组成。$[\mathbf{X},\mathbf{Y}]$ 代表 $\mathbf{X}$ 和 $\mathbf{Y}$ 在通道层上拼接起来。$\mathcal{P}_i^R$ 是一个基于拼接特征的具有 3×3 卷积核的预测输出层，输出一个单通道的残差图。一旦得到 $\mathbf{R}_i$ ，就可以通过 (公式1) 得到细化后的输出 $\mathbf{S}_i$。

为了更好地了解 BAM 的工作机制，在 图3，我们将 BAM 在不同层上的通道平均特征可视化。在 BAM 中，原始特征将首先分别乘以 FF 和 BF 注意力图而分为两个分支。直接相乘的结果显示在黄色（FF 特征）和蓝色（BF 特征）框的左半部分。我们可以看到 FF 分支将注意力转移到从其较高层预测出的前景区域，来探索前景显著性线索。在卷积层之后，对不确定区域给予更高的优先级。作为补充，BF 分支将重点放在背景区域上以探索背景线索，在其中寻找可能的显著对象。在我们的 BiANet 中, 自上而下的预测上采样是一个过程，在这个过程中显著对象的分辨率逐渐提高。这将导致不确定的粗略边界。提高对象边缘分割的质量对于分割任务很重要。许多基于主动轮廓 [51], [52], [54]–[57]，边缘监督 [42], [67], [89] 的方法，提出向边缘区域转移更多的注意力。这与我们的级联双边注意力模型的初衷和优势相吻合。 我们可以看到 FF 和 BF 特征都集中在这些边界上。低层和高分辨率 FF 分支将消除不确定区域的溢出，而 BF 分支将消除不属于背景的不确定区域。这就是 BiANet 在细节上表现更好并且易于预测锐利边缘的重要原因。联合推断后，我们可以看到双边增强的特征包含更多可区分的前景和背景空间信息。生成的残差分量在边缘处具有鲜明的对比度，从而抑制背景区域并增强前景区域。

C. BAM 的多尺度扩展（MBAM）

场景中显著对象的位置，大小和形状各不相同。因此，在高层中探索多尺度上下文有利于理解场景 [71], [88]。为此，我们将 BAM 扩展为多尺度版本，其中使用空洞卷积（dilated）组从不确定的前景和背景区域提取金字塔表示。具体来说，这个模块可以描述为

$$\mathbf{R}_i=\mathcal{P}_i^R\Big(\left[\sqcup_{j=1}^4\mathcal{D}_{ij}^F\left(\mathbf{F}_i\odot\mathbf{A}_i^F\right),\sqcup_{j=1}^4\mathcal{D}_{ij}^B\left(\mathbf{F}_i\odot\mathbf{A}_i^B\right)\right]\Big),\tag{6}$$

其中 $\sqcup$ 代表拼接操作。$\mathcal{D}_{i1}^F$ 和 $\mathcal{D}_{i1}^B$ 由一个 32 通道 1×1 卷积核和一个 ReLU 层组成。$\{\mathcal{D}_{ij}^F\}_{j=2}^4$ 和 $\{\mathcal{D}_{ij}^B\}_{j=2}^4$ 是空洞率为 3，5，7 的空洞卷积组。他们都由一个 32 通道的 3×3 卷积核和一个 ReLU 层组成。

我们建议在高层交叉模式特征中应用 MBAM，例如 $\{\mathbf{F}_3,\mathbf{F}_4,\mathbf{F}_5\}$，因为他们需要不同大小的接受域来探索多尺度背景。MBAM 有效地提高了检测性能，但也引入了一定的计算开销。因此, MBAM 的数量应在实际应用中进行权衡。在 节IV-C3中, 我们将详细讨论 MBAM 的数量如何改变检测效果和计算开销。

为了直观地观察 MBAM 带来的增益效果，我们在图4中对来自 MBAM 和 BAM 的平均前景优先特征图进行了可视化。在第二行中，特征图从模型直接获得，模型中三个 MBAM 位于其三层，而在最后一行，所有特征图都是从 BAM 获得的。我们可以看到目标对象（马）占场景的很大比例。由于无法有效地感知多尺度

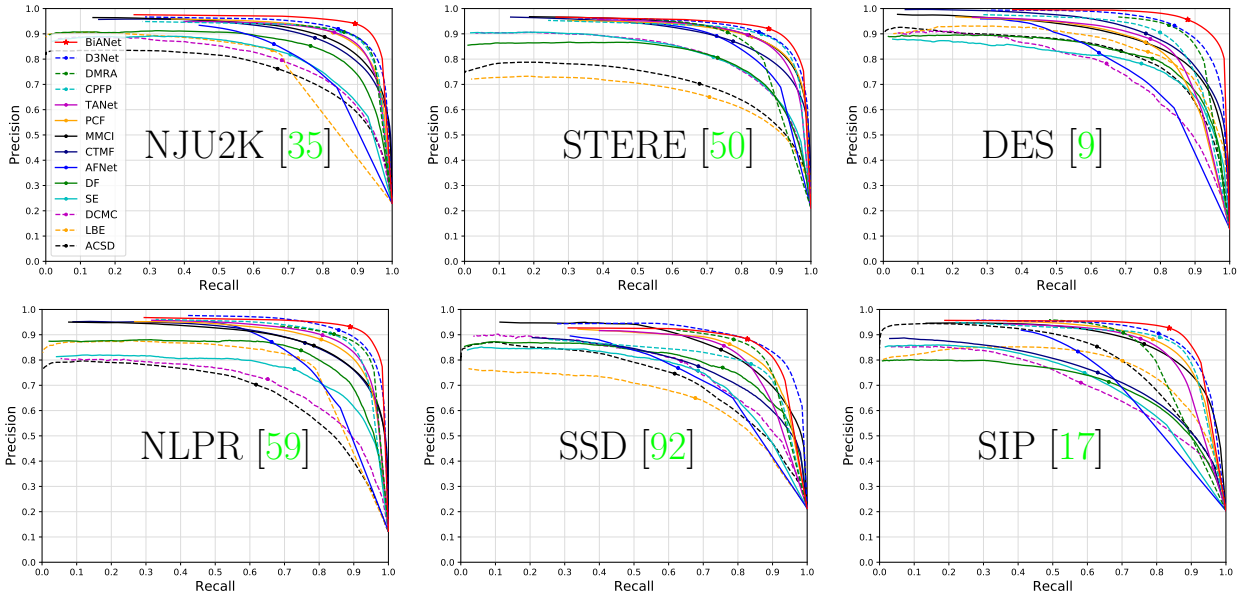

图 5. **我们的 BiANet 的 PR 曲线以及其他 6 种数据集上的 13 种最新方法**。每条曲线上的节点表示用于计算 maximal F-measure 的精确率和召回率。

信息，因此 BAM 无法在高水平上捕获准确的全局显着区域，最终导致不完整的预测。而当引入了多尺度扩展，我们可以看到更高级别的特征实现了更强的空间表示，支持定位更完整的显著对象。

D. 实现细节

1) 设定：在实现中我们对高层的边缘输出 $\{\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$ 应用 MBAM，同时在所有插值操作中使用双线性插值。我们 backbone 的初始参数是从 ImageNet 上经过预训练的 VGG-16 网络加载的。我们的 BiANet 基于 PyTorch [58].

2) 训练：延续 D3Net [17]，我们分别使用 NJU2K [35] 和 NLPR [59] 中的 1485、700 张图片对作为训练集。NJU2K 中剩下的样例用作验证集。We employ the Adam optimizer [37] with an initial learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. 我们使用 Adam [37] 作为优化器，学习率为 0.0001，且 $\beta_1 = 0.9$, $\beta_2 = 0.99$。batch size 设为 8。对于 VGG 作为 backbone，我们把 BiANet 训练 30 轮，对于 ResNet 和 Res2Net 作为 backbone，我们训练 50 轮。训练测试中，图片都被调整为大小：224×224。The output saliency maps are resized back to the original size for evaluation. 评估时，输出显著图被重新调整为原始大小。由一块 NVIDIA GeForce RTX 2080Ti 加速后，我们的 BiANet（VGG-16 为 backbone）训练时间少于三小时，在 224×224 分辨率下，BiANet 可以以 34∼80fps（视 MBAM 个数略有不同）速度运行。

IV. 实验

A. 评估机制

1) 评估数据集：我们在六个广泛使用的基于 RGB-D 的 SOD 数据集上进行了实验。NJU2K [35] 和 NLPR [59] 是两个分别包含 1985 张和 1000 张图片的大规模 RGB-D SOD 数据集，DES [9] 包含使用 Microsoft Kinect 收集的 135 张具有精细结构的室内图片 [85]。STERE [50] 包含 1000 个互联网图片，且相应的深度图是使用过滤流算法通过立体图像生成的 [45]. SSD [92] 是具有 400 张 960×1080 分辨率图片的小规模高分辨率数据集。SIP [17] 是具有 929 人像的高质量 RGB-D SOD 数据集。

2) 评估指标：我们采用九种指标来全面评估这些方法。Precision-Recall (PR) **曲线** [62] 展示了预测出的显著图在不同二进制阈值下的精度和召回性能。F-measure (F_β) [1] 通过阈值精度和查全率的加权谐波均值来计算。我们应用了 [4] 中提到的最大 F-measure。**平均绝对误差** (MAE, $\mathcal{M}$) [60] 直接估算预测值与二进制标签图之间的平均像素绝对差值。S-measure (S_α) [13] 是一个高级指标，它考虑了区域感知和对象感知的结构相似性。E-measure (E_ξ) [14] 是最近在二元图评估领域中提出的增强对齐方法，它在一项内将局部像素值与图

表 I

我们的 BiANet 和 9 个深度学习方法、四个传统方法，在六个主流数据集上的定量比较，比较项目分别是：S-measure (S_α), maximum F-measure (F_β), maximum E-measure (E_ξ), mean absolute error (MAE, $\mathcal{M}$), mean-square error (MSE), peak signal-to-noise ratio (PSNR), 和 structural similarity (SSIM)。↑ 代表数值越大模型越好，↓ 相反。对于传统方法，数据结果基于总体数据集而不是测试集。我们展示了不同 backbone 下 BiANet 的表现。vgg11 和 vgg16 是 [65] 中提出的 vgg 网络。Res50 是 [29] 中提出的 ResNet-50 网络。Res^{2}50 是 [26] 提出的 Res2Net-50 网络。比较中使用的是应用 VGG-16 为 backbone 的 BiANet。我们使用**加粗**标出最佳结果，使用 下划线 标出次佳结果。

		ACSD	LBE	DCMC	SE	DF	AFNet	CTMF	MMCI	PCF	TANet	CPFP	DMRA	D3Net	BiANet (Ours)			
	Metric	[35]	[21]	[12]	[27]	[63]	[70]	[28]	[8]	[6]	[7]	[87]	[61]	[17]	vgg16	vgg11	Res50	Res^{2}50
NJU2K [35]	$S_\alpha\uparrow$	0.699	0.695	0.686	0.664	0.763	0.772	0.849	0.858	0.877	0.878	0.879	0.886	0.893	0.915	0.912	0.917	0.923
	$F_\beta\uparrow$	0.711	0.748	0.715	0.748	0.804	0.775	0.845	0.852	0.872	0.874	0.877	0.886	0.887	0.920	0.913	0.920	0.925
	$E_\xi\uparrow$	0.803	0.803	0.799	0.813	0.864	0.853	0.913	0.915	0.924	0.925	0.926	0.927	0.930	0.948	0.947	0.949	0.952
	$\mathcal{M}\downarrow$	0.202	0.153	0.172	0.169	0.141	0.100	0.085	0.079	0.059	0.060	0.053	0.051	0.051	0.039	0.040	0.036	0.034
	MSE↓	0.105	0.117	0.106	0.127	0.079	0.087	0.045	0.044	0.039	0.041	0.041	0.043	0.035	0.030	0.030	0.029	0.027
	PSNR↑	10.76	11.13	11.09	10.84	12.67	12.55	14.75	15.20	16.44	16.33	16.60	16.93	17.22	18.96	18.71	19.14	19.48
	SSIM↑	0.336	0.811	0.512	0.691	0.546	0.822	0.689	0.699	0.822	0.832	0.891	0.903	0.866	0.913	0.909	0.923	0.926
STERE [50]	$S_\alpha\uparrow$	0.692	0.660	0.731	0.708	0.757	0.825	0.848	0.873	0.875	0.871	0.879	0.835	0.889	0.904	0.899	0.905	0.908
	$F_\beta\uparrow$	0.669	0.633	0.740	0.755	0.757	0.823	0.831	0.863	0.860	0.861	0.874	0.847	0.878	0.898	0.892	0.899	0.904
	$E_\xi\uparrow$	0.806	0.787	0.819	0.846	0.847	0.887	0.912	0.927	0.925	0.923	0.925	0.911	0.929	0.942	0.941	0.943	0.942
	$\mathcal{M}\downarrow$	0.200	0.250	0.148	0.143	0.141	0.075	0.086	0.068	0.064	0.060	0.051	0.066	0.054	0.043	0.045	0.040	0.039
	MSE↓	0.099	0.117	0.084	0.101	0.078	0.062	0.046	0.038	0.040	0.041	0.041	0.057	0.037	0.032	0.034	0.032	0.031
	PSNR↑	10.67	9.65	11.97	11.57	12.51	13.97	14.40	15.73	15.77	15.54	16.26	14.39	16.71	17.78	17.21	17.85	18.05
	SSIM↑	0.318	0.213	0.523	0.668	0.487	0.849	0.682	0.739	0.801	0.837	0.894	0.885	0.850	0.902	0.898	0.915	0.918
DES [9]	$S_\alpha\uparrow$	0.728	0.703	0.707	0.741	0.752	0.770	0.863	0.848	0.842	0.858	0.872	0.900	0.898	0.931	0.943	0.930	0.942
	$F_\beta\uparrow$	0.756	0.788	0.666	0.741	0.766	0.728	0.844	0.822	0.804	0.827	0.846	0.888	0.880	0.926	0.938	0.927	0.942
	$E_\xi\uparrow$	0.850	0.890	0.773	0.856	0.870	0.881	0.932	0.928	0.893	0.910	0.923	0.943	0.935	0.971	0.979	0.968	0.978
	$\mathcal{M}\downarrow$	0.169	0.208	0.111	0.090	0.093	0.068	0.055	0.065	0.049	0.046	0.038	0.030	0.033	0.021	0.019	0.021	0.017
	MSE↓	0.058	0.071	0.058	0.058	0.053	0.058	0.029	0.033	0.035	0.032	0.029	0.025	0.021	0.014	0.012	0.015	0.013
	PSNR↑	12.74	11.94	12.85	13.70	13.85	14.08	16.52	16.14	16.85	17.03	17.96	18.77	19.17	20.50	20.61	20.05	20.59
	SSIM↑	0.181	0.134	0.505	0.700	0.557	0.866	0.774	0.655	0.871	0.885	0.919	0.937	0.901	0.943	0.943	0.947	0.951
NLPR [59]	$S_\alpha\uparrow$	0.673	0.762	0.724	0.756	0.802	0.799	0.860	0.856	0.874	0.886	0.888	0.899	0.905	0.925	0.927	0.926	0.929
	$F_\beta\uparrow$	0.607	0.745	0.648	0.713	0.778	0.771	0.825	0.815	0.841	0.863	0.867	0.879	0.885	0.914	0.914	0.917	0.919
	$E_\xi\uparrow$	0.780	0.855	0.793	0.847	0.880	0.879	0.929	0.913	0.925	0.941	0.932	0.947	0.945	0.961	0.962	0.962	0.963
	$\mathcal{M}\downarrow$	0.179	0.081	0.117	0.091	0.085	0.058	0.056	0.059	0.044	0.041	0.036	0.031	0.033	0.025	0.024	0.023	0.023
	MSE↓	0.069	0.053	0.061	0.057	0.041	0.049	0.029	0.032	0.029	0.027	0.028	0.026	0.022	0.018	0.018	0.018	0.018
	PSNR↑	12.61	15.48	13.84	15.09	16.18	15.53	16.97	16.82	18.07	18.41	19.26	19.17	19.61	21.10	21.00	21.14	21.21
	SSIM↑	0.248	0.896	0.544	0.743	0.626	0.881	0.770	0.730	0.869	0.881	0.922	0.933	0.901	0.941	0.941	0.948	0.949
SSD [92]	$S_\alpha\uparrow$	0.675	0.621	0.704	0.675	0.747	0.714	0.776	0.813	0.841	0.839	0.807	0.857	0.865	0.867	0.861	0.863	0.863
	$F_\beta\uparrow$	0.682	0.619	0.711	0.710	0.735	0.687	0.729	0.781	0.807	0.810	0.766	0.844	0.846	0.849	0.839	0.843	0.843
	$E_\xi\uparrow$	0.785	0.736	0.786	0.800	0.828	0.807	0.865	0.882	0.894	0.897	0.852	0.906	0.907	0.916	0.899	0.911	0.901
	$\mathcal{M}\downarrow$	0.203	0.278	0.169	0.165	0.142	0.118	0.099	0.082	0.062	0.063	0.082	0.058	0.059	0.051	0.054	0.048	0.050
	MSE↓	0.107	0.138	0.102	0.128	0.089	0.104	0.066	0.049	0.042	0.044	0.069	0.050	0.040	0.040	0.043	0.040	0.042
	PSNR↑	10.61	9.44	11.61	11.18	12.55	12.01	13.22	14.84	16.22	15.94	14.96	15.95	16.68	17.72	17.34	17.49	17.62
	SSIM↑	0.257	0.195	0.491	0.663	0.542	0.811	0.706	0.732	0.846	0.850	0.861	0.900	0.865	0.902	0.894	0.914	0.911
SIP [17]	$S_\alpha\uparrow$	0.732	0.727	0.683	0.628	0.653	0.720	0.716	0.833	0.842	0.835	0.850	0.806	0.864	0.883	0.877	0.887	0.889
	$F_\beta\uparrow$	0.763	0.751	0.618	0.661	0.657	0.712	0.694	0.818	0.838	0.830	0.851	0.821	0.861	0.890	0.882	0.890	0.893
	$E_\xi\uparrow$	0.838	0.853	0.743	0.771	0.759	0.819	0.829	0.897	0.901	0.895	0.903	0.875	0.910	0.925	0.924	0.926	0.928
	$\mathcal{M}\downarrow$	0.172	0.200	0.186	0.164	0.185	0.118	0.139	0.086	0.071	0.075	0.064	0.085	0.063	0.052	0.054	0.047	0.047
	MSE↓	0.093	0.083	0.107	0.137	0.121	0.107	0.098	0.055	0.053	0.058	0.055	0.078	0.048	0.043	0.044	0.040	0.040
	PSNR↑	11.12	11.38	10.56	10.13	10.35	11.37	11.32	14.13	14.83	14.47	15.04	13.66	15.56	17.14	16.61	17.33	17.47
	SSIM↑	0.454	0.285	0.412	0.706	0.459	0.816	0.666	0.738	0.838	0.834	0.892	0.874	0.859	0.906	0.900	0.918	0.918

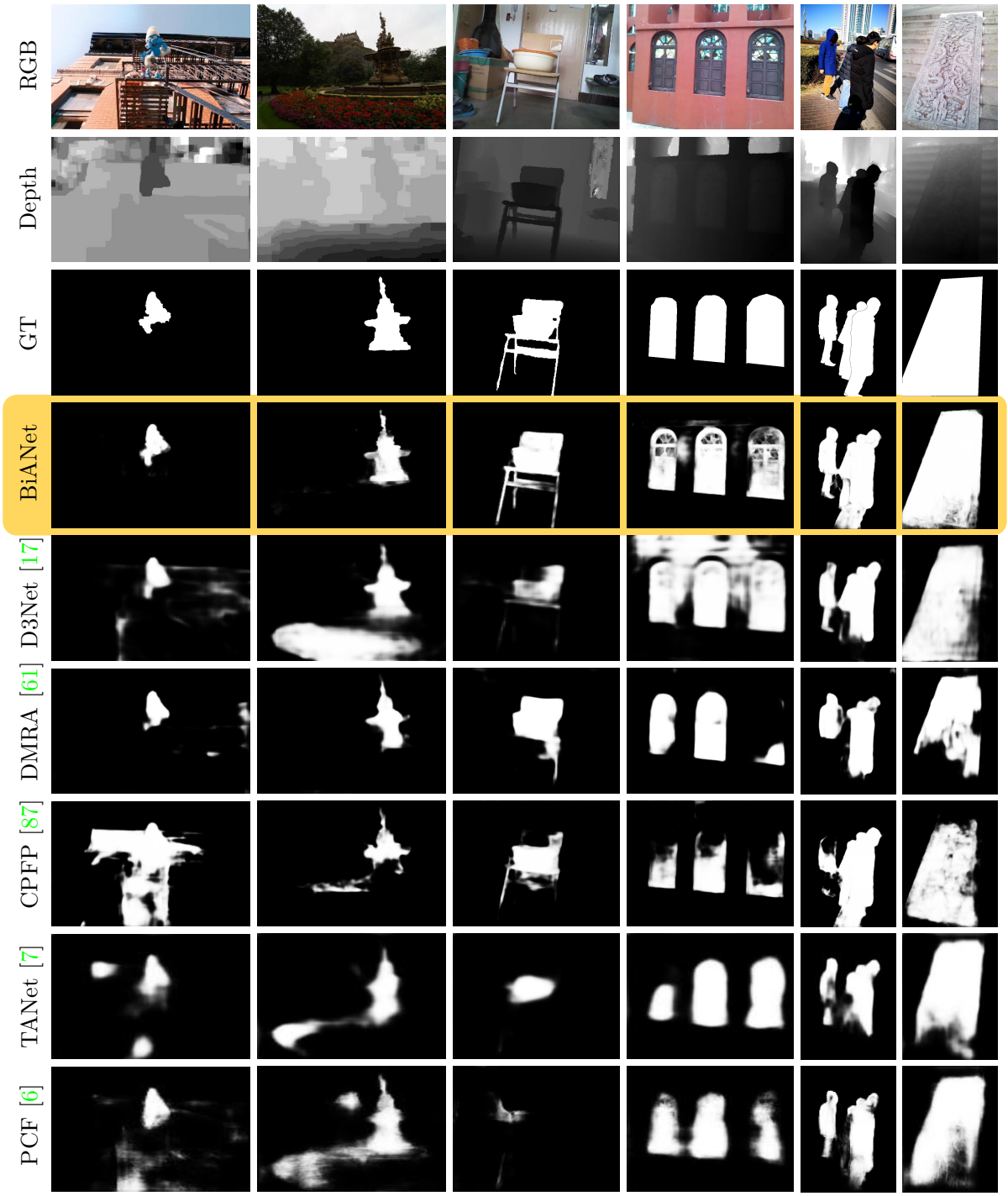

图 6. BiANet **和其他最好的五种方法的可视化对比**。输入包括不同的复杂场景，小物体（第一列），复杂背景（第一、二列），复杂结构（第三列），低对比度（第二、六列），低质量或复杂深度（第二、四、六列）和多物体（第四、五列）。

像级平均值结合在一起，从而共同捕获图像级统计信息和局部像素匹配信息。我们使用 [5], [40] 提出的最大 E-measure。 **均方误差** (MSE) 测量误差平方的平均值。**峰值信噪比** (PSNR) 是信号的最大可能功率与影响信号表示保真度的破坏噪声的功率之间的比率。PSNR 越高，预测质量越高。**结构相似度** (SSIM) 根据亮度，对比度和结构评估两个图像的相似性。MSE，PSNR，SSIM 这些评估方法在水印 [2], [3]，图片压缩 [72]，图片增强 [43] 中被广泛应用。

表 II
在 NJU2K 和 STERE 数据集上对**对提出结构的消融分析**。候选机制有深度信息（Dep），前景优先（FF），背景优先（BF）和多尺度扩展（ME）。ME 被应用在前三层特征中。

#	Candidates				NJU2K [35]		STERE [50]	
	Dep	FF	BF	ME	$F_\beta \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$
No. 1					0.881	0.885	0.882	0.893
No. 2	✓				0.903	0.904	0.887	0.894
No. 3	✓	✓			0.908	0.908	0.895	0.901
No. 4	✓		✓		0.910	0.908	0.892	0.900
No. 5	✓	✓	✓		0.915	0.913	0.897	0.903
No. 6	✓	✓		✓	0.913	0.911	0.900	0.904
No. 7	✓		✓	✓	0.912	0.911	0.893	0.902
No. 8		✓	✓	✓	0.905	0.903	0.894	0.901
No. 9	✓	✓	✓	✓	0.920	0.915	0.898	0.904

B. 与最新方法对比

1) Comparison methods: 我们对比了 13 个 STOA RGB-D 方法, 包括四种传统方法: ACSD [35], LBE [21], DCMC [12], MDSF [66], 和 SE [27]，和 9 个基于深度神经网络的方法：DF [63], AFNet [70], CTMF [28], MMCI [8], PCF [6], TANet [7], CPFP [87], DMRA [61], 和 D3Net [17]. 这些方法的代码和显著图由作者提供。

2) 定量评估: 完整的定量评估结果列于 表I。根据这些指标的综合性能，从右到左介绍了参与比较的方法，其中 MAE($\mathcal{M}$) 的值越低，模型的效果越好。其他指标则相反。我们还绘制了这些方法的 PR 曲线 图5。可以看到，我们的 BiANet 与比较方法相比具有明显的优势。这些数据集中，DMRA [61] 和 D3Net [17] 配合比较好。在大型数据集 NJU2K [35] 和 NLPR [59] 上，我们的 BiANet 在 F_β 上提高了 ∼3%，超过了第二好的记录。在 DES [9] 数据集上，相较于严重依赖深度信息的其他方法，我们提出的 BiANet 也在 F_β 上有 3.8% 的提升。这表明我们的 BiANet 可以更有效地利用深度信息。尽管 SSD [92] 数据集是高分辨率的，但是深度图的质量很差。我们的 BiANet 始终超出专门为低质量深度图设计的高鲁棒性方法 D3Net [17]。同时我们的 BiANet 在包含复杂场景、多物体的数据集 SIP [17] 上表现最好。

3) 定性结果: 为了进一步证明我们的 BiANet 的有效性，我们将 BiANet 的显着性图和其他五种方法的可视化结果可视化在 图6。可以看到，第一列中的目标对象很小，白色的鞋子和帽子很难与背景区分开。我们的 BiANet 有效地利用了深度信息，而其他信息则受到 RGB 背景杂波的干扰。第二列中的输入更具有挑战性，因为深度图标签错误，并且 RGB 图像是在黑暗的环境中以低对比度拍摄的。我们的 BiANet 成功地检测到目标雕塑并消除了花朵和雕塑基部的干扰，而 D3Net 错误地检测到了更接近的玫瑰花结，而 DMRA 丢失了与背景相似的物体部分。第三列显示了我们的 BiANet 能够检测复杂结构和无纹理 [53] 显著物体的能力。在这些方法中，只有我们的 BiANet 才能完全发现椅子，包括细腿。第四列是一个多对象场景。由于下面的三个显著窗口与墙之间没有深度差异，因此它们不会在深度图上反映出来，但是在深度图上可以清楚地观察到上面的三个窗口。在这种情况下，深度图将误导后续的分割。我们的 BiANet 可以从 RGB 图像中检测出多个物体，而噪声却更少。第五列也是一个多对象场景。深度图的下半部分与地面干扰相混淆。由于对不确定区域进行了自上而下的细化，BiANet 可以检测到大部分细节，例如腿部区域和中间位置的人。最后一列是一个大型对象，其颜色和深度图无法区分。大规模，低色彩对比度以及缺乏可辨别的深度信息，使场景非常具有挑战性。幸运的是，我们的 BiANet 在这种情况下具有非常强的鲁棒性。

C. 消融研究与分析

在本节中，我们主要研究：1) 双边注意力机制对我们的 BiANet 的好处；2) BAM 在我们的 BiANet 上对 RGB-D SOD 具有不同程度的有效性；3) 在我们的 BiANet 的不同级别上对 MBAM 的进一步改进；4) 将 BAM 和 MBAM 结合在一起用于 RGB-D SOD 的好处；5) 不同主干对 BiANet 的 RGB-D SOD 的影响；6) 检测非最前端物体的鲁棒性。

1) 双边注意力机制的有效性: 我们对 NJU2K [35] 和 STERE [50] 数据集进行消融研究，以研究所提出方法中不同机制的贡献。这两个数据集包含大规模样本和各种场景。因此，对这两个数据集进行评估可以更好地反映不同设置的性能。 此处使用的基准模型包含 VGG-16 主干和精炼残差结构 (residual refine structure)。它以 RGB 图像作为输入，没有深度信息。在没有任何其他机制的情况下，我们的

表 III

在每个边缘输出中添加或删除 BAM/MBAM 对准确性的影响。 'None' 不含任何 BAM/MBAM 的基准模型. 也就是说, 'None' 是 表II的 No.2。'w/ L*i*' 代表在 'None' 的基础上给第 i 级添加 BAM/MBAM。'All' 是每层都有 BAM/MBAM 的基准模型。'w/o L*i*' 代表在 'All' 基准模型上删掉第 i 级上的 BAM/MBAM。

	Metric	w/ L1	w/ L2	w/ L3	w/ L4	w/ L5	None	w/o L1	w/o L2	w/o L3	w/o L4	w/o L5	All
BAM	$S_\alpha\uparrow$	0.908	0.909	0.908	0.906	0.904	0.904	0.911	0.911	0.913	0.912	0.913	0.913
	$F_\beta\uparrow$	0.910	0.911	0.909	0.905	0.904	0.903	0.914	0.914	0.915	0.915	0.915	0.915
	$E_\xi\uparrow$	0.944	0.945	0.943	0.943	0.941	0.942	0.945	0.948	0.947	0.947	0.948	0.948
	$\mathcal{M}\downarrow$	0.043	0.043	0.044	0.044	0.045	0.046	0.041	0.041	0.041	0.042	0.041	0.041
MBAM	$S_\alpha\uparrow$	0.908	0.909	0.910	0.910	0.910	0.904	0.916	0.916	0.914	0.913	0.912	0.916
	$F_\beta\uparrow$	0.909	0.912	0.909	0.911	0.911	0.903	0.920	0.918	0.916	0.914	0.913	0.919
	$E_\xi\uparrow$	0.944	0.945	0.945	0.946	0.947	0.942	0.951	0.947	0.948	0.945	0.946	0.948
	$\mathcal{M}\downarrow$	0.044	0.043	0.042	0.042	0.042	0.046	0.038	0.039	0.039	0.040	0.040	0.039

表 IV

MBAM **的准确性和计算成本分析**. $\times0\sim\times5$ 代表从高层适配到低层 MBAM 的个数。fps 代表每秒帧数。Params 代表参数大小。FLOPs = 浮点运算参数。精准度 F_β 和 $\mathcal{M}$ 从 NJU2K 数据集中测算得出。计算开销 fps 和 FLOPs 是在 224×224 分辨率上进行测试的。Train 代表训练时间。 注意, $\times3$ 即为节IV-B提到的标准设定。

	$\times0$	$\times1$	$\times2$	$\times3$	$\times4$	$\times5$	D3Net [17]	DMRA [61]
$F_\beta\uparrow$	0.914	0.917	0.918	0.920	0.920	0.919	0.887	0.886
$\mathcal{M}\downarrow$	0.041	0.040	0.040	0.039	0.038	0.039	0.051	0.051
fps↑	∼80	∼65	∼55	∼50	∼42	∼34	∼55	∼40
Params ↓	45.0M	46.9M	48.7M	49.6M	50.1M	50.4M	145.9M	59.7M
FLOPs ↓	34.4G	35.0G	36.2G	39.1G	45.2G	58.4G	55.7G	121.0G
Train ↓	0.58h	0.66h	0.81h	1.05h	1.49h	2.29h	-	-

基本网络的性能如下所示: 表II No. 1, 基于网络, 我们逐渐添加不同的机制并测试各种组合。这些候选机制有深度信息（Dep）, 前景优先注意力（FF）, 背景优先注意力（BF）和多尺度扩展（ME）。表II No. 3, 通过应用 FF, 性能得到了一定程度的提高。通过将注意力转移到前景对象上, 可以有效地学习前景提示, 从而从中受益。仅使用 BF 时, 我们获得了相似的精度, 如 No. 4 所示。它擅长区分背景中的显著区域和非显著区域, 并有助于在不确定的背景下找到显著物体的更完整区域; 但是, 过多的注意力集中在背景上, 而对前景的线索却没有很好的探索, 这就导致有时会引入背景噪声。

当我们将 FF 与 BF 结合在一起, 形成 BAM 并将其应用于所有边缘输出时, 性能将得到提高。我们可以看到, 与 No. 2 相比, BAM 的 S 度量增加了 0.9%, 最大 F 度量增加了 1.2%。当我们将前三级 BAM 替换为 MBAM 时, 性能进一步提高。

为了进一步验证同时挖掘前景和背景线索的重要性, 我们在保持其他组件不变的情况下去除背景优先或前景优先注意力, 并将结果记录在 No. 6 和 No. 7 中。我们可以看到, 如果没有前景优先或背景优先的注意力, 所提出模型的性能将会降低。此外, 深度信息对于双边注意力也很重要。它提供了丰富的前景和背景关系。我们删除了 No. 8 中预测精度受损的深度信息。

2) *不同层次的 BAM 的效能*: 为了验证我们的 BAM 模块在每个功能级别上是否有效, 我们将 BAM 分别应用于 No. 2 模型特征提取器的每一侧输出。也就是说, 在每个实验中, BAM 被应用于一侧的输出, 而其他的则经过几个卷积组, 而没有被前景优先/背景优

先注意图增强。具体来说，为了替换 BAM，通道递减特征经过两个卷积组，每个卷积组由一个具有 32 个内核的卷积层和一个 ReLU 层组成。然后，单个卷积层跟随这两个组来预测残差。从 表III, 我们可以看到，每一层中的 BAM 都有助于检测性能的改进。此外，我们发现在较低级别应用的 BAM 对结果的贡献更大。为了进一步证实我们的观察，我们在所有五个侧输出特征中应用 BAM 作为基线模型。之后, 我们删除了五个 BAM 中的一个，性能显示在 表III中。我们可以看到，从低级特征中移除 BAM 会导致性能损失。

3) 不同级别上 MBAM 的效能: 在 表II中，与 5 号相比，9 号在其更高的三个层次 $\{\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$ 上进行多尺度延伸。这种扩展有效地提高了模型的性能。为了更好地展示 MBAM 在每一层特征中的增益，类似于上一节，我们分别将 MBAM 应用于 2 号模型的每一侧输出。实验结果记录在表III中。同样，我们也在所有五个边缘输出特征中应用 MBAM 作为基准模型。然后，我们删除五个 MBAM 之一以观察性能损失。 可以看出，不同级别的 MBAM 对结果带来不同程度的提升。比较 BAM 和 MBAM，我们可以看到一个更有趣的现象，低层应用 BAM 带来更多提升，而应用在更高级别的 MBAM 更有效。

4) BAM 与 MBAM 的合作: 上述观察指导我们，在合作使用 BAM 和 MBAM 时，我们应该优先考虑更高层次的 BAM 的多尺度扩展。因此，我们从上到下扩展 BAM，直到所有 BAM 都转换为 MBAM。我们在表IV中记录了逐步扩展过程中的最终检测性能和计算成本。我们从最高层开始，并逐渐将 MBAM 的数量增加到三个。我们可以看到对模型的效果是稳步提升的，但是计算成本也增加了。在较低的层里，添加 MBAM 没有明显的效果。这一现象符合我们的预期。此外，由于分辨率高，下层 BAM 的扩展会增加计算成本并降低健壮性。MBAM 数量的选择需要平衡应用场景的精度和速度要求。对速度要求较高的场景，不建议使用 MBAM。我们最轻量级的模型可以达到 ~80fps，同时确保显著的性能优势。

参数大小和 FLOPs 优于 SOTA 方法 D3Net [17] 和 DMRA [61]。在需要高精度的场景中，我们建议在更高级别的特征上应用不超过三个 MBAM。

5) 不同主干网下的表现: 我们基于其他一些广泛使用的主干网来实现 BiANet 证明所提出的双边注意机制对不同特征提取器的有效性。具体来说，除了 VGG-16 [65]，我们提供了 BiANet 在 VGG-11 [65]、ResNet-50 [29] 和 Res2Net-50 [26] 上的结果。与 VGG-16 相比，VGG-11 是更轻量级的主干。如 表I所示，虽然精度略低于 VGG-16，它仍然以更快的速度到达 SOTA。具有更强主干的 BiANet 将带来更显著的改进。例如，当我们使用 ResNet-50 (如 D3Net [17]) 作为主干网，与 D3Net [17] 相比，我们的 BiANet 在 MAE 方面对 NJU2K [35] 带来了 1.5% 的改进。当配备 Res2Net-50 [26] 时，与 SOTA 方法相比，BiANet 在 maxF 方面在 NJU2K [35] 上实现了 3.8% 的改进。

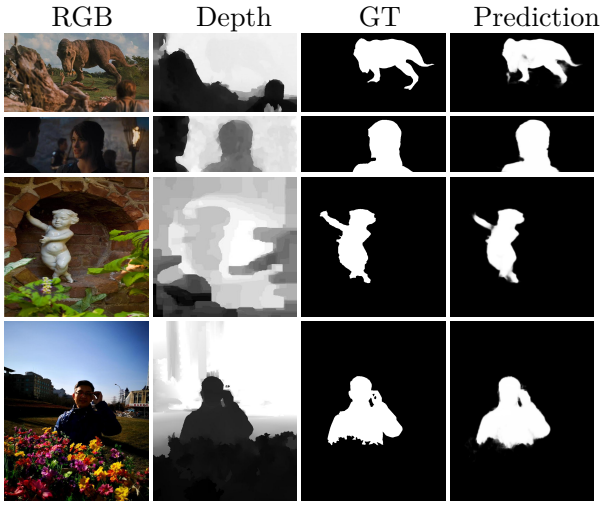

图 7. **使用 BiANet 检测非最前面的显著对象**。显著的物体在岩石后面（见第一排的恐龙），而不是最前面的人（第二排），或者在花的后面（最后两排）。我们可以看到 BiANet 的预测对于非最前面的显著对象检测问题是鲁棒的。

++

6) 检测非最前面物体的鲁棒性: 在实际应用中，我们的 BiANet 不要求显著对象位于场景的最前面。BiANet 联合探索来自 RGB 图像和深度图的显著性线索。当深度图带来距离模糊时，我们的 BiANet 在大多数情况下仍然可以很好地依赖其他线索，例如中心性、深度图中的形状以及来自 RGB 信息的丰富线索。图7中的示例证明了我们的 BiANet 在处理此类场景时的鲁棒性。

D. 失败案例分析

在 图8中，我们举例说明了 BiANet 在某些极端环境中工作时的一些失败案例。BiANet 利用深度信息提供的关系在前景和背景区域双边探索显著性线索。当深度图带来距离模糊时，我们的 BiANet 在大多数情况下仍然是健壮的，这取决于其他线索，例如中心性、深度图中的形状和来自 RGB 信息的丰富线索等。然而，图8

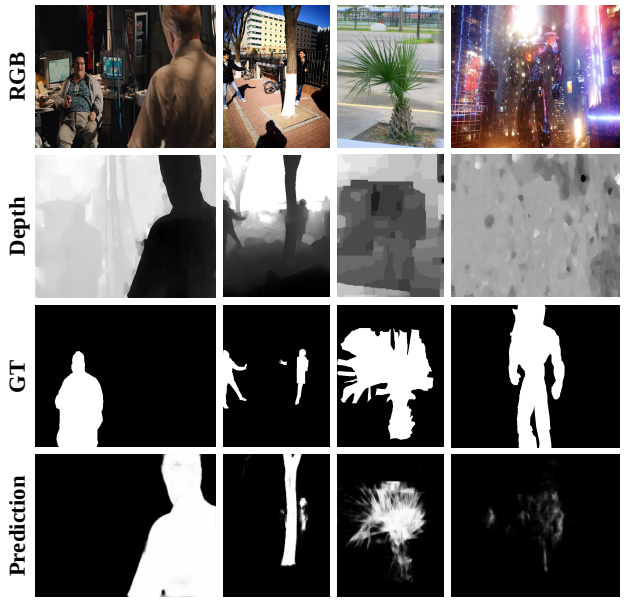

图 8. BiANet **在极端环境下的失败案例**。在前两列中，由于靠近观察者的对象不是目标，因此深度图提供了误导性信息。在最后两列中，BiANet 因混乱的 RGB 信息和粗略的深度图而失败。

中的前两列是极端的例子。具体来说，我们可以看到目标对象在 RGB 图像和深度图中都存在混淆。

另一种可能导致失败的情况是 BiANet 在复杂场景中遇到粗深度图时（见最后两列）。在第三列中，深度图提供了不准确的空间信息，影响了细节的检测。在最后一列中，不准确的深度图和混乱的 RGB 信息使 BiANet 无法定位目标对象。

V. 总结

在本文中，我们为 RGB-D 显著性对象检测 (SOD) 任务提出了一种快速而有效的双边注意网络 (BiANet)。为了更好地利用前景和背景信息，我们提出了一个双边注意模块（BAM）来囊括前景优先注意和背景优先注意机制。为了充分利用多尺度技术，我们将 BAM 模块扩展到其多尺度版本 (MBAM)，以捕获更好的全局信息。在六个基准数据集上进行的大量实验表明，得益于 BAM 和 MBAM 模块，我们的 BiANet 在定量和定性性能方面优于以前在 RGB-D SOD 上的最先进方法。我们提出的 BiANet 在单个 GPU 上以实时速度运行，使其成为各种实际应用程序的潜在解决方案。

参考文献

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In IEEE CVPR, pages 1597–1604, 2009.

[2] Deepayan Bhowmik and Charith Abhayaratne. Quality scalability aware watermarking for visual content. IEEE Trans. Image Process., 25(11):5158–5172, 2016.

[3] Deepayan Bhowmik and Charith Abhayaratne. Embedding distortion analysis in wavelet-domain watermarking. ACM Trans. on Multimedia Comput., Commun., Appl., 15(4):1–24, 2019.

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. IEEE Trans. Image Process., 24(12):5706–5722, 2015.

[5] Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion. IEEE Trans. Image Process., 29:4296–4307, 2020.

[6] Hao Chen and Youfu Li. Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection. In IEEE CVPR, pages 3051–3060, 2018.

[7] Hao Chen and Youfu Li. Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. Image Process., 28(6):2825–2835, 2019.

[8] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognition, 86:376–385, 2019.

[9] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In ICIMCS, page 23, 2014.

[10] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. IEEE Trans. Circuits Syst. Video Technol., 2018.

[11] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from rgb to rgbd saliency: A depth-guided transformation model. IEEE Trans. Cybern., pages 1–13, 2019.

[12] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Process. Lett., 23(6):819–823, 2016.

[13] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In IEEE ICCV, pages 4548–4557, 2017.

[14] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In IJCAI, pages 698–704, 2018.

[15] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. arXiv preprint arXiv:2007.03380, 2020.

[16] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In IEEE CVPR, pages 2919–2929, 2020.

[17] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Trans. Neural Netw. Learn. Syst., 2021.

[18] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In IEEE CVPR, pages 8554–8564, 2019.

[19] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: Rgb-d salient object detection with a bifurcated backbone strategy network. In ECCV, 2020.

[20] Xingxing Fan, Zhi Liu, and Guangling Sun. Salient region detection for stereoscopic images. In IEEE DSP, pages 454–458, 2014.

[21] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for RGB-D salient object detection. In IEEE CVPR, pages 2343–2350, 2016.

[22] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In IEEE CVPR, pages 1623–1632, 2019.

[23] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In IEEE CVPR, pages 3052–3062, 2020.

[24] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In IEEE CVPR, pages 3052–3062, 2020.

[25] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. arXiv preprint arXiv:2008.12134, 2020.

[26] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[27] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for rgb-d image via saliency evolution. In IEEE ICME, pages 1–6, 2016.

[28] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Trans. Cybern., 28(6):2825–2835, 2018.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE CVPR, pages 770–778, 2016.

[30] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell., 41(4):815–828, 2019.

[31] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In NeurIPS, 2018.

[32] Laurent Itti and Christof Koch. Computational modelling of visual attention. Nature reviews neuroscience, 2(3):194–203, 2001.

[33] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell., 20(11):1254–1259, 1998.

[34] Peng Jiang, Zhiyi Pan, Changhe Tu, Nuno Vasconcelos, Baoquan Chen, and Jingliang Peng. Super diffusion for salient object detection. IEEE Trans. Image Process., 2019.

[35] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In IEEE ICIP, pages 1115–1119, 2014.

[36] Chanho Jung and Changick Kim. A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. IEEE Trans. Image Process., 21(3):1272–1283, 2011.

[37] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[38] Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan S. Kankanhalli, and Shuicheng Yan. Depth matters: influence of depth cues on visual saliency. In ECCV, pages 101–115, 2012.

[39] Guanbin Li, Yukang Gan, Hejun Wu, Nong Xiao, and Liang Lin. Cross-modal attentional context learning for rgb-d object detection. IEEE Trans. Image Process., 28(4):1591–1601, 2019.

[40] Gongyang Li, Zhi Liu, and Haibin Ling. ICNet: Information conversion network for rgb-d based salient object detection. IEEE Trans. Image Process., 29:4873–4884, 2020.

[41] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. GradNet: Gradient-guided network for visual object tracking. In IEEE ICCV, 2019.

[42] Xiangtai Li, Xia Li, Li Zhang, Cheng Guangliang, Jianping Shi, Zhouchen Lin, Yunhai Tong, and Shaohua Tan. Improving semantic segmentation via decoupled body and edge supervision. In ECCV, 2020.

[43] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In IEEE CVPR, pages 3867–3876, 2019.

[44] Fangfang Liang, Lijuan Duan, Wei Ma, Yuanhua Qiao, Zhi Cai, and Laiyun Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing, 275:2227–2238, 2018.

[45] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell., 33(5):978–994, 2011.

[46] Guanghai Liu and Dengping Fan. A model of visual attention for natural image retrieval. In IEEE ICISCC, pages 728–733. IEEE, 2013.

[47] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell., 33(2):353–367, 2010.

[48] Yi Liu, Jungong Han, Qiang Zhang, and Caifeng Shan. Deep salient object detection with contextual information guidance. IEEE Trans. Image Process., 29:360–374, 2019.

[49] V. Mahadevan and N. Vasconcelos. Saliency-based discriminant tracking. In IEEE CVPR, pages 1007–1013, 2009.

[50] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In IEEE CVPR, pages 454–461, 2012.

[51] Joanna Isabelle Olszewska. Active contour based automatic feedback for optical character recognition. In BIOSIGNALS, pages 318–324, 2014.

[52] Joanna Isabelle Olszewska. Active contour based optical character recognition for automated scene understanding. Neurocomputing, 161:65–71, 2015.

[53] Joanna Isabelle Olszewska. Where is my cup? - fully automatic detection and recognition of textureless objects in real-world images. In Computer Analysis of Images and Patterns, pages 501–512, Cham, 2015. Springer International Publishing.

[54] Joanna Isabelle Olszewska, Christophe De Vleeschouwer, and Benoit Macq. Speeded up gradient vector flow b-spline active contours for robust and real-time tracking. In IEEE ICASSP, volume 1, pages I–905. IEEE, 2007.

[55] Joanna Isabelle Olszewska, Christophe De Vleeschouwer, and Benoit Macq. Multi-feature vector flow for active contour tracking. In IEEE ICASSP. IEEE, 2008.

[56] Joanna Isabelle Olszewska, Tom Mathes, Christophe De Vleeschouwer, Justus Piater, and Benoit Macq. Non-rigid object tracker based on a robust combination of parametric active

contour and point distribution model. In VCIP, volume 6508, page 65082A. International Society for Optics and Photonics, 2007.

[57] Joanna I Olszewska and Thomas L McCluskey. Ontology-coupled active contours for dynamic video scene understanding. In IEEE ICIES, pages 369–374. IEEE, 2011.

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In NeurIPS, pages 8024–8035, 2019.

[59] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In ECCV, pages 92–109, 2014.

[60] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In IEEE CVPR, pages 733–740, 2012.

[61] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In IEEE ICCV, pages 7254–7263, 2019.

[62] David M W Powers. Evaluation: from Precision, Recall and F-measure to ROC, informedness, markedness and correlation. J. of Mach. Learn. Technol., 2(1):37–63, 2011.

[63] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. RGBD salient object detection via deep fusion. IEEE Trans. Image Process., 26(5):2274–2285, 2017.

[64] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for rgb-d saliency detection. In IEEE CVPR Workshop, pages 25–32, 2015.

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[66] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Trans. Image Process., 26(9):4204–4216, 2017.

[67] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In ECCV, 2020.

[68] Chung-Chi Tsai, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection and co-segmentation via progressive joint optimization. IEEE Trans. Image Process., 28(1):56–71, 2018.

[69] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. Artificial Intelligence, 78(1-2):507–545, 1995.

[70] Ningning Wang and Xiaojin Gong. Adaptive fusion for RGB-D salient object detection. CoRR, abs/1901.01369, 2019.

[71] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven C. H. Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In IEEE CVPR, pages 1448–1457, 2019.

[72] Yanxiang Wang, Charith Abhayaratne, Rajitha Weerakkody, and Marta Mrak. Colour space transforms for improved video compression. In IWSSIP 2014 Proceedings, pages 219–222. IEEE, 2014.

[73] Yupei Wang, Xin Zhao, Xuecai Hu, Yin Li, and Kaiqi Huang. Focal boundary guided salient object detection. IEEE Trans. Image Process., 28(6):2813–2824, 2019.

[74] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In IEEE ICCV, 2019.

[75] Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, and Yu Zhang. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In IEEE CVPR, 2017.

[76] Xiaolin Xiao, Yicong Zhou, and Yue-Jiao Gong. Rgb-'d' saliency detection with pseudo depth. IEEE Trans. Image Process., 28(5):2126–2139, 2018.

[77] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In IEEE CVPR, pages 3166–3173, 2013.

[78] Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. Bifurcated backbone strategy for rgb-d salient object detection. arXiv e-prints, pages arXiv–2007, 2020.

[79] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In IEEE CVPR, 2020.

[80] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. Uncertainty inspired rgb-d saliency detection. arXiv preprint arXiv:2009.03075, 2020.

[81] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In IEEE CVPR, pages 8582–8591, 2020.

[82] Lihe Zhang, Jie Wu, Tiantian Wang, Ali Borji, Guohua Wei, and Huchuan Lu. A multistage refinement network for salient object detection. IEEE Trans. Image Process., 29:3534–3545, 2020.

[83] Pingping Zhang, Wei Liu, Huchuan Lu, and Chunhua Shen. Salient object detection with lossless feature reflection and weighted structural loss. IEEE Trans. Image Process., 28(6):3048–3060, 2019.

[84] Wei Zhang and Hantao Liu. Study of saliency in objective video quality assessment. IEEE Trans. Image Process., 26(3):1275–1288, 2017.

[85] Zhengyou Zhang. Microsoft kinect sensor and its effect. IEEE Trans. Multimedia, 19(2):4–10, 2012.

[86] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In ECCV, pages 455–472, 2020.

[87] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In IEEE CVPR, pages 3927–3936, 2019.

[88] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In IEEE CVPR, pages 3085–3094, 2019.

[89] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multi-class part parsing with joint boundary-semantic awareness. In IEEE ICCV, 2019.

[90] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. Computational Visual Media, 2020.

[91] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In IEEE ICME, 2019.

[92] Chunbiao Zhu and Ge Li. A Three-pathway Psychobiological Framework of Salient Object Detection Using Stereoscopic Technology. In IEEE ICCV Workshop, pages 3008–3014, 2017.

张钊 是南开大学媒体计算实验室硕士研究生，师从程明明教授。在此之前，他于 2019 年获得扬州大学学士学位。他的研究兴趣主要集中在低级视觉，如显著物体检测、交互式分割和图像增强。

林铮 现为南开大学计算机学院博士生，师从程明明教授。他的研究兴趣包括深度学习、计算机图形学和计算机视觉。

徐君 分别于 2011 年和 2014 年在南开大学数学科学学院获得信息与概率学士学位，硕士学位，并于 2014 年获得香港理工大学博士学位。他曾在阿拉伯联合酋长国的起源人工智能研究院担任助理研究员。

金闻达 现为天津大学硕士研究生，师从郭伟教授。在此之前，他是安徽大学的一名本科生。他的研究兴趣主要为计算机视觉，显著性目标检测。

卢少平 现为南开大学计算机学院副教授。博士毕业于清华大学。他的研究兴趣主要集中在视觉计算的交叉领域，特别关注 3D 图像和视频处理、计算摄影和表示、视觉场景分析和机器学习。

范登平 于 2019 年在南开大学获得博士学位。并于当年加入起源人工智能研究院。发表 CVPR、ICCV、ECCV 等顶级期刊和会议论文约 25 篇。他的研究兴趣包括计算机视觉和视觉注意，特别是在 RGB 显着物体检测 (SOD)、RGB-D SOD、视频 SOD、Co-SOD。他在 IEEE CVPR 2019 上获得了最佳论文决赛入围奖，和 IEEE CVPR 2020 最佳论文奖提名。