基于级联互信息最小化的 RGB-D 显著性检测

张静 [1] 范登平 [2,⋆] 戴玉超 [3] 于昕 [4] 钟怡然 [5] Nick Barnes[1] 邵岭 [2]
[1] 澳洲国立大学 [2] 起源人工智能研究院（IIAI） [3] 西北工业大学 [4] 悉尼科技大学 [5] 商汤科技

摘要

现有的 *RGB-D* 显著性检测模型没有显式利用 *RGB* 和深度数据以实现有效地多模态学习。本文提出了新颖的多阶段级联学习框架，通过互信息最小化**显式地**建模 *RGB* 图像和深度数据之间的多模态信息。特别地，本文首先将各模态的特征映射为低维特征向量，然后采用互信息最小化正则项来减少 *RGB* 图像特征和深度数据所提供的几何特征之间的冗余。随后执行多阶段级联学习来将互信息最小化约束加入网络的每个阶段中。*RGB-D* 显著性基准数据集上的大量实验验证了该框架的有效性。

此外，为了促进该领域发展，本文贡献了最大的（超过 *NJU2K* 七倍多）*COME15K* 数据集，包含 *15,625* 个带有高质量多边形/笔画/目标/实例/排名等级的标注图像对。基于这些丰富的标签，本文额外构造了四个新的基准，并设计了高性能的基准模型。本文还观测到一些有趣的现象，这些现象将促进未来的模型设计。源代码和数据集可见：*https://github.com/JingZhang617/cascaded_rgbd_sod*。

1. 引言

显著性检测模型被用来发现图像中吸引人注意的区域。传统的显著性检测主要应用于 RGB 图 [37, 44, 22, 43, 36]。随着深度数据的普及 (表 1)，RGB-D 显著性检测 [46, 35, 52, 54] 获得了极大的关注。深度数据提供了真实的几何信息，对图像前景与背景相似的情景尤为重要。此外，深度感知器 (如：Microsoft Kinect) 对光线变化的鲁棒性也有利于显著性检测任务。

因为 RGB 和深度数据捕获同一场景下的不同信息，现有的 RGB-D 显著性检测模型 [35, 1, 3, 2, 54, 50, 30, 12, 19, 33, 52] 主要用不同的融合策略隐式建模 RGB 和深度数据之间的补充关系。三种主要的融合策略被广泛研究：早融合 [38, 46]、晚融合 [41, 16, 36] 与跨层级融合 [35, 1, 3, 2, 54, 50, 30, 12, 19, 33, 52, 25]。虽然有效地融合策略能提升性能，但没有在网络设计上加约束去强迫它学习模态之间的互补信息，并且我们也不能显式地评价深度数据在这些模型中的贡献 [53]。

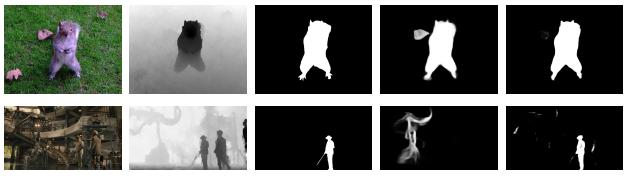


Image Depth GT BBSNet Ours

图 1. 目前最优 RGB-D 显著性检测模型，例如: BBSNet [12] 和本文模型的显著性预测结果比较。

作为多模态学习任务，一个训练好的模型应该在网络能力内最大化不同模态之间的联合熵。最大化联合熵也等价于最小化互信息，可以阻止网络关注冗余信息。为了显式地建模 RGB 图像和深度数据之间的互补信息，本文通过互信息最小化设计一个多阶段级联学习框架。特别地，本文把互信息最小化作为一个正则项（如图 2所示），这样带来两个好处：1）显式地建模图像特征和几何特征之间的冗余；2）在互信息最小化的约束下有效地融合图像特征和几何特征。图 1所示结果证明了本方法生成的显著性图像的有效性。

此外，我们发现目前没有大规模的 RGB-D 显著性检测训练集。在表 1中，本文分别从数据集规模、数据类型、深度数据的来源以及其在显著性检测中的角色（用于训练“Tr”或用于测试“Te”）方面比较了广泛使用的 RGB-D 数据集。本文注意到传统的 RGB-D 显著

⋆ 通讯作者：范登平 (*dengpfan@gmail.com*)。
本工作是张静在 IIAI 实习期间，由范登平研究员指导下完成的。
本文为 ICCV2021 论文 [47] 的中文翻译版。

表 1. 和广泛使用的 RGB-D 数据集的比较。

Dataset	Size	Type	Depth Source	Role
NJU2K[21]	1,985	Movie/Internet	FujiW3 camera + optical flow	Tr, Te
DUT [35]	1,200	Indoor/Outdoor	Light-field cameras	Tr, Te
NLPR [34]	1,000	Indoor/Outdoor	Microsoft Kinect	Tr, Te
SSB [32]	1,000	Internet	Stereo cameras	Te
SIP [11]	929	Person in outside	Huawei Mate10	Te
DES [7]	135	Indoor	Microsoft Kinect	Te
LFSD [27]	80	Indoor/Outdoor	Lytro Illum cameras	Te
Ours	**15,625**	Indoor/Outdoor	Holopix Social Platform	Tr, Te

性检测的训练集是来自 NJU2K [21] 和 NLPR [34] 数据集的样本组合，一共仅有 2200 个图像对。虽然另外 800 张来自 DUT 数据集 [35] 的训练图像也可以作为训练集的第三种数据源，但这也总共仅有 3000 副图像。因此，数据量不够大可能会导致学习的模型有偏差。此外，我们观察到在现有的 RGB-D 训练集中存在相似背景的图像，比如，超过 10% 的训练数据来自相同的场景且有相似的光照条件。这种缺乏多样性的数据集可能导致模型的泛化能力较差。同时我们也注意到最大的测试集 [11] 只包含 1000 个图像对，这也不足以充分评价 RGB-D 显著性检测深度学习模型的综合性能。

为了构建一个用于训练鲁棒性的模型的数据集，以及一个足够大的测试数据集用于模型评价，本文贡献了最大的 RGB-D 显著性检测数据集。本文从 Holo50K 数据集 [18] 采集了 8025 个图像对用于训练以及 7600 个图像对用于测试。我们不仅提供了二值型标注，而且有用于立体显著性检测的标注，还有用于弱监督 RGB-D 显著性检测的笔画标注和多边形标注，以及实例级的 RGB-D 显著性标注和 RGB-D 显著性排序。此外，本文贡献了 5000 个无标签的训练图像用于半监督或自监督的 RGB-D 显著性检测。

本文主要贡献有：1）设计了用于 RGB-D 显著性检测的多阶段级联学习框架，通过互信息最小化“显式地”建模 RGB 图像和深度信息之间的冗余。2）本文的互信息最小化正则项可以方便地拓展到其他多模态学习流程中，以建模多模信息中的冗余。3）贡献了最大的 RGB-D 显著性检测数据集，共有 15625 个有标签数据和 5000 个无标签数据以实现全监督/弱监督/无监督的 RGB-D 显著性检测。4）构建了新的 RGB-D 显著性检测评测基准，并且引入了基准模型用于立体的和弱监督的 RGB-D 显著性检测。

2. 相关工作

2.1. RGB-D 显著性检测模型

对于 RGB-D 显著性检测，一个主要的焦点就是探索 RGB 图像和深度数据之间的互补信息。前者提供了关于某一场景的图像信息，后者引入了几何信息。根据来自两个模态的信息如何被融合的方式，当前的 RGB-D 显著性检测模型可以被划分为三种类别：早融合模型 [38, 46]、晚融合模型 [41, 16, 36]、跨层级融合模型 [35, 1, 3, 2, 54, 50, 30, 12, 19, 33, 52, 25, 26]。第一类方法直接拼接 RGB 图像和其深度，而晚融合模型分别处理每个模态（RGB 和深度），然后在输出层进行融合。上述两个方法在输入层或者输出层执行多模态融合，而跨层级融合模型在特征空间进行融合。具体来说，RGB 图像和深度数据的特征在网络的不同层 [33, 26, 12, 25, 31, 5, 49, 30, 50] 被逐级融合。虽然当前的模型融合了 RGB 图像和深度数据用于多模态学习，但是都没有显式地说明网络是如何实现有效地多模态学习。本文提出了一个如图 2所示的跨层级的融合模型。通过设计“互信息正则项”，从而减少图像特征和几何特征间的冗余以进行有效地多模态学习。

2.2. 在 RGB-D 数据集上多模态学习

多模态学习的基本假设是各个模态内都存在共同的信息和特有的信息。对于 RGB-D 数据集，RGB 图像和深度数据共享相似的语义信息，这可以定义为共同的部分。RGB 图像编码图像信息，包括像素强度或者对象的颜色。而深度数据编码了几何信息，给出目标对象的相对几何位置。图像信息和几何信息之间的不同点恰是这两个模态的多样性的展现。对于 RGB-D 数据实现多模态学习的主要关注点就是使用不同的融合策略 [42, 4, 31, 24]，例如早融合、晚融合、或者跨层级融合。不同于传统的解决方案，本文设计了多阶段级联学习框架，通过互信息最小化减少各个模态间特征的冗余。虽然互信息最大化 [29, 39] 被广泛用于表示学习中，来产生一个与输入相似的表示。但本文采用互信息最小化作为正则项，来减少特征冗余，以进行有效地多模态学习。

2.3. RGB-D 显著性数据集

表 1列出了广泛使用的 RGB-D 显著性检测数据集，包括 NJU2K[21]、NLPR [34]、SSB[32]、DES[7]、LFSD[27]、SIP [11]、DUT[35]。典型的训练数据集包括

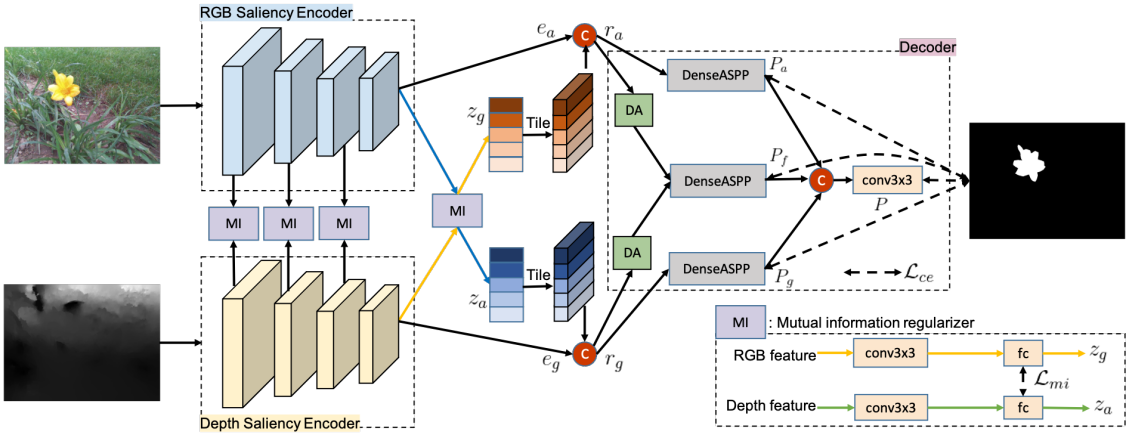


图 2. 本文提出的用于 RGB-D 显著性检测的多阶段级联学习框架图。本文将 RGB 图像和深度数据输入显著性编码器中，以提取各个模态的显著性特征，同时用互信息正则项迫使从两个模态中学到不同的特征。然后融合各模态（z_a 和 z_g）的低维度特征和原始图像特征（e_a 和 e_g）以有效地建模两个模态的互补信息，并获得最终预测 P。“DenseASPP”模块是来自 [45] 的稠密空洞空间金字塔池化模块，“DA”是来自 [14] 的双注意力模块。

了来自 NJU2K[21] 的 1485 副图片以及来自 NLPR[34] 的 700 副图片。最近，Piao 等人 [35] 引入了 DUT 数据集，有 800 副图像用于训练和 400 副图像用于测试。为了促进 RGB-D 显著性检测任务发展，本文引入了最大的 RGB-D 显著性检测训练集和测试集，具体在章节4中介绍。

3. 本文的 CMINet

如图 2所示，本文引入了多阶段级联学习框架用来显式地建模 RGB-D 显著性检测的互补信息。

3.1. 显著性编码器

本文将训练集表示为 $T = \{X_i, Y_i\}_{i=1}^{N}$，其中 i 为图像索引，N 为训练集的大小，X_i 和 Y_i 分别是输入 RGB-D 图像对和其相应的显著性图像真值（GT）。本文将训练图像对数据（RGB 图像 I 和深度 D）传到显著性编码器中，如图 2所示，分别用来提取图像特征 $f_{\alpha_a}(I)$ 和几何特征 $f_{\alpha_g}(D)$，其中，α_a 和 α_g 分别表示 RGB 显著性编码器和深度显著性编码器的参数。

本文基于 ResNet50 网络 [17] 构建了显著性编码器，包括四个卷积阶段 $\{s^1, s^2, s^3, s^4\}$。本文在每个 $s^c \in \{s^c\}_{c=1}^{4}$ 后加入额外的一层核尺寸为 3×3 的卷积层将 s^c 的输出通道维度减少至 $C = 32$，从而获得特征图 $\{e^1, e^2, e^3, e^4\}$。RGB 显著性编码模块的最终输出为 $e_a = \{e_a^1, e_a^2, e_a^3, e_a^4\}$，深度显著性编码器的输出为 $e_g = \{e_g^1, e_g^2, e_g^3, e_g^4\}$。要指出的是，RGB 显著性编码器和深度显著性编码器共享相同网络结构但不共享权重。

3.2. 特征嵌入

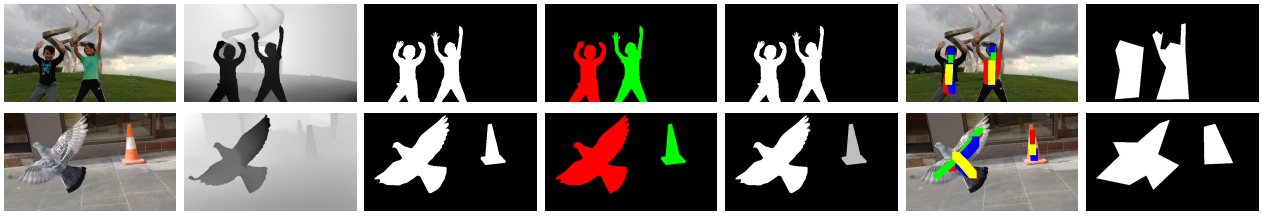
给定 RGB 显著性编码器的输出 $e_a = \{e_a^1, e_a^2, e_a^3, e_a^4\}$ 和深度显著性编码器的输出 $e_g = \{e_g^1, e_g^2, e_g^3, e_g^4\}$，目标是将 RGB 特征和深度特征映射到低维的特征空间中以实现特征嵌入。特别地，本文提出了一个多阶段级联学习策略可以在网络中每个阶段执行互补信息学习。对于靠前阶段，本文将 RGB 特征 $\{e_a^c\}_{c=1}^{3}$ 和深度特征 $\{e_g^c\}_{c=1}^{3}$ 传给两个不同的 3×3 卷积层（图 2中的“conv3x3”）分别获得 RGB 分支和深度分支下通道数为 $4 * C$ 的特征图。

然后本文采用两个全连接层（图 2中的“fc”）将通道数 $4 * C$ 的特征图映射为两个不同的、维度为 $K = 6$ 的低维特征向量 $\{z_a^c\}_{c=1}^{3}$ 和 $\{z_g^c\}_{c=1}^{3}$。该互补学习相关的损失函数（在章节3.3和3.5中介绍）被用于减少靠前阶段中 RGB 图和深度图之间的特征冗余。在最后阶段，本文首先在空间维度中扩展 (tile) 低维的特征向量 z_a^4 和 z_g^4，然后将其和另一模态的原始图像特征拼接，从而获得 $4 * C + K$ 通道数的 RGB 分支和深度分支的特征图 r_a 和 r_g。

3.3. 多模态学习

在获取到关于 RGB 图像和深度数据的特征嵌入 z_a 和 z_g 后，本文引入互信息最小化正则项以显式地减

本文将拼接的 $\{e_a^c\}_{c=1}^{4}$ 和 $\{e_g^c\}_{c=1}^{4}$ 分别代表原始 RGB 特征和原始深度特征。



(a) (b) (c) (d) (e) (f) (g)

图 3. 本文的新 RGB-D 显著性检测数据集的标注: (a) RGB 图像、(b) 深度数据、(c) 二值型标注、(d) 实例级标注、(e) 基于排序的标注、(f) 笔画标注、(g) 多边形标注。本文的多样性标注将促进开发不同的全/弱监督 RGB-D 显著性检测模型。

少两个模态之间的冗余。本文的基本假设是一个好的图像显著性特征和几何显著性特征对应该都含有共同的部分（语义相关）和不同的部分（领域相关）。互信息 M_I 被用于衡量熵之间的不同：

$$M_I(z_a, z_g) = H(z_a) + H(z_g) - H(z_a, z_g), \quad (1)$$

其中 $H(.)$ 表示熵，$H(z_a)$ 和 $H(z_g)$ 是边际熵，$H(z_a, z_g)$ 是 z_a 和 z_g 的联合熵。直观上说，有两个潜在变量（或者条件熵）的 KL 散度（KL）如下：

$$KL(z_a||z_g) = H_{z_g}(z_a) - H(z_a), \quad (2)$$

$$KL(z_g||z_a) = H_{z_a}(z_g) - H(z_g), \quad (3)$$

其中 $H_{z_g}(z_a) = -\sum_x z_a(x) \log z_g(x)$ 是交叉熵。然后将等式 1， 2， 3相加得到：

$$\begin{aligned} M_I(z_a, z_g) =& H_{z_g}(z_a) + H_{z_a}(z_g) - \\ & H(z_a, z_g) - (KL(z_a||z_g) + KL(z_g||z_a)). \end{aligned} \quad (4)$$

给定 RGB 图像和深度数据，$H(z_a, z_g)$ 是非负的，从而最小化互信息可以通过最小化：$\mathcal{L}_{mi} = (H_{z_g}(z_a) + H_{z_a}(z_g)) - (KL(z_a||z_g) + KL(z_g||z_a))$ 实现。由于，$M_I(z_a, z_g)$ 衡量了给定 z_g 的观测数据时 z_a 不确定性的减少程度或者反之亦然。作为一个多模态学习任务，每个模态应该从其他模态学到一些任务相关的新特性。通过最小化 $M_I(z_a, z_g)$，本文可以有效地探索各模态间的互补信息。注意到，虽然 [46] 中使用 KL 损失项作为概率分布的相似度衡量标准，但本文使用它来衡量多模态学习中模态间的相似度。

3.4. 显著性解码器

有了互信息作为正则项，本文在网络的靠前阶段约束了冗余的特征，并且在最后阶段获得了优化的 RGB 显著性特征 r_a 以及优化的深度显著性特征 r_g。在 r_a 之后，本文采用一个 DenseASPP [45] 模块来获得带有多级上下文信息的 RGB 显著性预测 P_a。同样的，本文可以获得深度显著性预测 P_g。显著性解码器 f_γ（图 2中的“Decoder”）接收优化的显著性特征 r_a，r_g 以及 RGB 显著性预测 P_a 和深度显著性预测 P_g 作为输入，产生最终的预测 P，其中 γ 是该显著性解码器的参数集合。特别地，本文在 r_a 和 r_g 之后加了位置注意力模块 [14] 和通道注意力模块来分别突出特有特征来获得 $da(r_a)$ 和 $da(r_g)$。然后本文拼接 $da(r_a)$ 和 $da(r_g)$ 并输入给 DenseASPP [45] 模块以获得显著性预测结果 P_f。为了进一步融合模块之间的信息，本文按照通道拼接 P_a，P_g 和 P_f 并且传入 3×3 的卷积层来获得最终预测 P。

3.5. 目标函数

本文采用了二值交叉熵损失 $\mathcal{L}_{ce}$ 作为目标函数来训练多阶段级联学习框架，其中等式 (1)所示的互补约束，强迫 RGB 图像的显著性特征的分布不同于深度数据的显著性特征。最终的目标函数为：

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{ce}(P, Y) + \lambda_1 \mathcal{L}_{ce}(P_f, Y) + \lambda_2 \mathcal{L}_{ce}(P_a, Y) + \\ & \lambda_3 \mathcal{L}_{ce}(P_g, Y) + \lambda \sum_{c=1}^{4} \mathcal{L}_{mi}(z_a^c, z_g^c), \end{aligned} \quad (5)$$

根据经验，设置 $\lambda_1 = 0.8, \lambda_2 = 0.6, \lambda_3 = 0.4$。因为 $\mathcal{L}_{mi}$ 的取值范围是 $\mathcal{L}_{ce}$ 的十倍大，本模型设置其损失权重 $\lambda = 0.1$ 用于平衡训练。

4. COME15K 数据集

如表1所示，现有的 RGB-D 显著性检测训练集数据量不足，可能导致模型泛化能力差。此外，因为训练集是来自 NJU2K[21] 和 NLPR 数据集 [34] 的样本组合，训练集不同的划分常常导致不一致的模型评价结

果。最后，小数据量的测试集也不能充分评价 RGB-D 显著性检测模型的表现。为促进 RGB-D 显著性检测领域的研究，本文贡献了最大的 RGB-D 显著性检测数据集。本文提供了如图 3所示的二值型标注、实例级标注、基于排名的标注以及弱标注等。有关数据集的详细分析可见补充材料。

4.1. 数据集标注

本文新构建的 COME15K 数据集基于立体数据集 Holo50K[18]，其中包含室内和室外的场景。本文首先过滤Holo50K 数据集，获取到 16000 个立体图像对用于标注（候选的标注集合）以及另外 5000 个图像对作为无标签集合。注意到 Holo50K 中的立体图像对是直接通过双目照相机捕获且未经过矫正，所以本文使用一个改进的、目前最优的、现成的立体匹配算法 [56]，为候选标注集合和无标签集合计算深度数据，其中的输入是左右的视角图像。

为了给候选标注集合提供标注，我们首先让五个“粗粒度”标注者采用笔画标注方法标注每副图像中的显著性区域（只使用右视角图像）。第二步，“细粒度”的标注者将会划分显著性目标的整体范围，然后提供实例层级的标注。第三步，执行“大多数投票”来获取二值型显著性 GT 图，用于 RGB-D 显著性检测任务。注意，我们删除了那些没有公共显著性区域的样本，并且获取到最终的数据量为 15,625 的有标签数据集。此外，基于笔画标注和实例层级的显著性图像，本文根据初始的笔画标注，对每个显著性实例排序，形成了 RGB-D 显著性排序数据集。

本文也提供了弱标注用于弱监督 RGB-D 显著性检测，包括笔画标注和多边形标注。本文将多个粗粒度标注者的笔画标注中的多数原则作为数据集的笔画标注结果。特别地，本文首先获取了对应笔画标注中的多数派的实例，然后我们定义在该多数派对应的实例下的笔画标注作为数据集的笔画标注。最后，本文对该多数派对应的显著性实例进行多边形标注形成基于多边形的标注结果。

4.2. 数据集划分

本文将有标签数据集分为有 8,025 个样本的训练集和两个不同的分别有 4,600 和 3,000 个数据的测试集，称为“普通”和“困难”集。8,025 副训练集图像通过从有标签集中任意选择产生。对于测试集，本文引入两套不同困难程度的集合。特别地，本文将 RGB 图像基于全局和内部的对比度进行排序，并将带有低全局对比度的和高内部对比度的样本作为困难样本。从而得到 1,800 个困难样本 D_d 和 5,800 个普通样本 D_n 作为候选。本文随机从 D_d 中选择 30% 的样本以及 D_n 中选择 70% 的样本以获取“普通”测试集，剩余的作为“困难”测试集。

5. 实验

本文比较了我们的方法 CMINet 和现有的 RGB-D 显著性检测模型，并且展示了如表2 & 3所示的性能。此外，我们在新的训练集上重新训练目前最佳的 RGB-D 显著性检测模型，并且在表6中提供了这些模型在本文测试集上的表现。

5.1. 实验设定

数据集： 为了和现有 RGB-D 显著性检测模型公平比较，本文遵循传统的训练设置，其中训练集为来自 NJU2K[21] 的 1,485 条数据和来自 NLPR[34] 的 700 条数据的组合。然后我们测试了模型的表现，并在 NJU2K，NLP，LFSD[27]，DES[7]，SSB[32]，SIP[11] 和 DUT[35] 等测试集上进行比较。

评价指标： 本文在四个黄金评价指标上评价模型表现，即平均绝对误差（$\mathcal{M}$），平均 F 度量（F_β），平均 E 度量（E_ξ）[10] 和 S 度量（S_α）[9]，这部分在补充材料中进一步阐述。

训练细节： 我们的模型用 *Pytorch* 库实现。两个显著性编码器共享相同的网络结构，使用在 ImageNet 预训练的 ResNet50[17]，其余新加入的层被任意初始化。我们重新调整所有输入图像和真值的尺寸到 352×352 像素。设置最大的训练轮数（epoch）为 100，并且学习率初始值为 5e-5。采用 “step” 学习率衰减策略，衰减周期为 80 步，衰减率为 0.1。在 NVIDIA GeForce RTX 2080Ti GPU 上训练，batch size 为 5，传统训练集 (NJU2K-train+NLPR-train) 训练需要 4.5 小时，新训练集 (COME15K-train) 需要 16 小时。

5.2. 模型比较

定量比较： 本文比较了我们的 CMINet 和目前最先进的 RGB-D 显著性检测模型的表现，如表2所示。注意

删除了暴力图像。

关于图片全局和内部对比度的细节在补充材料中介绍。

表 2. 三个领先的基于手工设计特征的模型和十八个深度模型 (*) 在六个 RGB-D 显著性数据集上的基准结果。↑ & ↓ 分别表示越大或越小是越好。这里本文采用平均 F_β 和平均 E_ξ [10]。

		Early Fusion Models				Late Fusion Models						Cross-level Fusion Models											
	Metric	DF	DANet	UCNet	JLDCF	LHM	DESM	CDB	A2dele	AFNet	CTMF	DMRA	PCF	MMCI	TANet	CPFP	S2MA	BBS-Net	CoNet	HDFNet	BiaNet	CMWNet	CMINet
		[38]*	[55]*	[46]*	[15]*	[34]	[7]	[28]	[36]*	[41]*	[16]*	[35]*	[1]*	[3]*	[2]*	[54]*	[30]*	[12]*	[19]*	[33]*	[52]*	[25]*	Ours*
NJU2K	$S_\alpha \uparrow$	.763	.897	.897	.902	.514	.665	.632	.873	.822	.849	.886	.877	.858	.879	.878	.894	.921	.911	.908	.915	.903	**.939**
	$F_\beta \uparrow$	.653	.877	.886	.885	.328	.550	.498	.867	.827	.779	.873	.840	.793	.841	.850	.865	.902	.903	.892	.903	.881	**.925**
	$E_\xi \uparrow$	.700	.926	.930	.935	.447	.590	.572	.913	.867	.846	.920	.895	.851	.895	.910	.914	.938	.944	.936	.934	.923	**.956**
	$\mathcal{M} \downarrow$	.140	.046	.043	.041	.205	.283	.199	.051	.077	.085	.051	.059	.079	.061	.053	.053	.035	.036	.038	.039	.046	**.032**
SSB	$S_\alpha \uparrow$	.757	.892	.903	.903	.562	.642	.615	.876	.825	.848	.835	.875	.873	.871	.879	.890	.908	.896	.900	.904	.905	**.921**
	$F_\beta \uparrow$	.617	.857	.884	.873	.378	.519	.489	.874	.806	.758	.837	.818	.813	.828	.841	.853	.883	.877	.870	.879	.872	**.895**
	$E_\xi \uparrow$	.692	.915	.938	.936	.484	.579	.561	.925	.872	.841	.879	.887	.873	.893	.911	.914	.928	.939	.931	.926	.928	**.959**
	$\mathcal{M} \downarrow$	.141	.048	.039	.040	.172	.295	.166	.044	.075	.086	.066	.064	.068	.060	.051	.051	.041	.040	.041	.043	.043	**.034**
DES	$S_\alpha \uparrow$	.752	.905	.934	.931	.578	.622	.645	.881	.770	.863	.900	.842	.848	.858	.872	.941	.933	.906	.926	.931	.934	**.953**
	$F_\beta \uparrow$	.604	.848	.919	.907	.345	.483	.502	.868	.713	.756	.873	.765	.735	.790	.824	.909	.910	.880	.910	.910	.909	**.926**
	$E_\xi \uparrow$	.684	.961	.967	.959	.477	.566	.572	.913	.809	.826	.933	.838	.825	.863	.888	.952	.949	.939	.957	.948	.955	**.970**
	$\mathcal{M} \downarrow$	.093	.028	.019	.021	.114	.299	.100	.030	.068	.055	.030	.049	.065	.046	.038	.021	.021	.026	.021	.021	.022	**.015**
NLPR	$S_\alpha \uparrow$	.806	.908	.920	.925	.630	.572	.632	.887	.799	.860	.899	.874	.856	.886	.888	.916	.930	.900	.923	.925	.917	**.941**
	$F_\beta \uparrow$	.664	.850	.891	.894	.427	.430	.421	.871	.755	.740	.865	.802	.737	.819	.840	.873	.896	.859	.894	.894	.877	**.909**
	$E_\xi \uparrow$	.757	.945	.951	.955	.560	.542	.567	.933	.851	.840	.940	.887	.841	.902	.918	.937	.950	.937	.955	.948	.939	**.964**
	$\mathcal{M} \downarrow$	.079	.031	.025	.022	.108	.312	.108	.031	.058	.056	.031	.044	.059	.041	.036	.030	.023	.030	.023	.024	.029	**.019**
LFSD	$S_\alpha \uparrow$	.791	.845	.864	.862	.557	.722	.520	.831	.738	.796	.847	.794	.787	.801	.828	.837	.864	.842	.854	.845	.876	**.877**
	$F_\beta \uparrow$	.679	.826	.855	.848	.396	.612	.376	.829	.736	.756	.845	.761	.722	.771	.811	.806	.843	.834	.835	.834	.862	**.862**
	$E_\xi \uparrow$	.725	.872	.901	.894	.491	.638	.465	.872	.796	.810	.893	.818	.775	.821	.863	.855	.883	.886	.883	.871	.900	**.911**
	$\mathcal{M} \downarrow$	.138	.082	.066	.070	.211	.248	.218	.076	.134	.119	.075	.112	.132	.111	.088	.094	.072	.077	.077	.085	.066	**.064**
SIP	$S_\alpha \uparrow$	.653	.878	.875	.880	.511	.616	.557	.826	.720	.716	.806	.842	.833	.835	.850	.872	.879	.868	.886	.883	.867	**.894**
	$F_\beta \uparrow$	.465	.829	.867	.873	.287	.496	.341	.827	.702	.608	.811	.814	.771	.803	.821	.854	.868	.855	.875	.873	.851	**.887**
	$E_\xi \uparrow$	.565	.914	.914	.918	.437	.564	.455	.887	.793	.704	.844	.878	.845	.870	.893	.905	.906	.915	.923	.913	.900	**.933**
	$\mathcal{M} \downarrow$	.185	.054	.051	.049	.184	.298	.192	.070	.118	.139	.085	.071	.086	.075	.064	.057	.055	.054	.047	.052	.062	**.044**

表 3. 在 DUT [35] 测试集上的模型表现。

Metric	UCNet	JLDCF	A2dele	DMRA	CPFP	S2MA	CoNet	HDFNet	CMINet
	[46]	[15]	[36]	[35]	[54]	[30]	[19]	[33]	Ours
$S_\alpha \uparrow$	.907	.905	.884	.886	.749	.903	.919	.905	**.928**
$F_\beta \uparrow$	.902	.884	.889	.883	.695	.881	.911	.889	**.921**
$E_\xi \uparrow$	.931	.932	.924	.924	.759	.926	.947	.929	**.959**
$\mathcal{M} \downarrow$	.038	.043	.043	.048	.100	.044	.033	.040	**.030**

到，本文和目前的 RGB-D 显著性检测模型设置一样使用 NJU2K 和 NLPR 组成的训练集。本文模型一致更优的测试结果表明了本方法的有效性。此外，本文观测到当前 RGB-D 显著性检测模型间性能的差异是非常不易察觉的，如 BBS-Net [12]，CoNet[19]，HDFNet[33]，BiaNet[52]，和 CMWNet [52]，这表明了对于更大、更多元的训练和测试集用于模型训练和评价的必要性。

DUT [35] 数据集上的表现： 一些现有的 RGB-D 显著性检测方法 [35, 30] 在 DUT 训练集 [35] 上微调他们的模型来评价他们在 DUT 测试集上的性能。为了在 DUT 测试集上测试本文模型，我们遵守了相同的训练策略。在表3中，所有的模型在传统的训练集上训练，然后在 DUT 训练集上微调。一致的更优性能表明了本文模型的优越性。此外，因为表3中模型目前的测试结果是通过训练-重训练的方式实现（在联合训练集上训练，然后在 DUT 训练集 [35] 上重训），本文在传统训练集和 DUT 训练集的组合上重训了这些模型，然后观测到一致更差的表现。这个观测表明上述三个训练集中（NJU2K，NLPR 和 DUT）可能存在不一致的标注。这激励我们去收集更大的带有一致标注的训练集用于鲁棒的模型训练。

定性比较： 本文进一步在图 1中可视化了我们的预测结果。定性比较证明了采用本文提出的学习策略，本文的模型可以有效探索两个模态进行多模态学习。更多的结果展示在补充材料中。

模型尺寸和运行时间： 本文的模型大小为 84M，与目前最优的模型相当，如 BBS-Net[12] 的模型大小为 100M。本文模型实现每秒 10 张图片的推断，同样也与目前最优的模型相当。

5.3. 消融实验

本文执行了下列的消融实验来进一步分析模型的组件。本文也实现了没采用提出策略的基准模型用于突出互信息最小化正则项的贡献。注意到，所有的这些实验都在传统数据集上训练。

基准模型的性能： 为了测试本文设计的图2 所示的编码器和解码器的性能，本文在框架中删除了“互信息正则项”部分，并且直接拼接 RGB 特征 e_a 和深度特征 e_g 然后将其输入到解码器。该性能被标为表4中的“基准”。我们观测到“基准”模型与现有的 RGB-D 显著性检测模型有相似的性能。相比于本文最终的结果，“基准”模型更差的性能表明了提出的使用互信息作为正则项用于冗余约束这一方法的优越性。

详细的结构被展示在补充材料中。

表 4. 额外实验的性能结果。

	NJU2K[21]				SSB [32]				DES [7]				NLPR [34]				LFSD [27]				SIP [11]			
Method	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
Base	.910	.900	.935	.035	.890	.870	.917	.043	.926	.915	.959	.018	.920	.898	.942	.024	.842	.835	.880	.077	.879	.876	.917	.049
K3	.928	.908	.947	**.032**	.909	.892	.939	.036	.934	.922	.964	.018	.925	.904	.956	.022	.869	.845	.898	.067	.885	.879	.919	.047
K32	.924	.909	.944	.033	.908	.894	.941	.036	.938	.923	.966	.017	.927	.906	.959	.021	.856	.853	.900	.065	.885	.878	.921	.046
SS	.926	.913	.943	.034	.914	.882	.942	.036	.946	.927	.968	.017	.932	.896	.954	.021	.861	.852	.896	.067	.885	.879	.925	.046
W0	.918	.907	.944	.033	.892	.877	.923	.042	.934	.924	.964	.017	.924	.900	.945	.023	.843	.836	.881	.076	.884	.878	.916	.048
W1	.919	.909	.946	**.032**	.905	.886	.937	.037	.938	.927	.971	.016	.923	.903	.956	.022	.857	.853	.891	.071	.887	.882	.921	.045
P_f	.925	.908	.945	.033	.908	.887	.939	.036	.946	.925	.965	.016	.938	.907	.962	.023	.862	.845	.896	.068	.889	.886	.927	.045
S_{rgb}	.898	.890	.930	.040	.899	.876	.924	.042	.891	.883	.920	.028	.908	.885	.932	.031	.817	.807	.853	.095	.860	.865	.905	.056
S_{rgbd}	.915	.901	.932	.037	.903	.878	.931	.039	.920	.908	.942	.021	.914	.893	.943	.026	.850	.841	.886	.071	.876	.870	.910	.051
CMINet	**.939**	**.925**	**956**	**.032**	**.921**	**.895**	**.959**	**034**	**.953**	**.926**	**.970**	**.015**	**.941**	**.909**	**.964**	**.019**	**.877**	**.860**	**.911**	**.064**	**.894**	**.887**	**.933**	**.044**

特征空间的维度： 本文设置低维特征空间嵌入（z_a 和 z_g）的维度为 $K = 6$。为了测试特征维度对于网络性能的影响，我们设置 $K = 3$ 和 $K = 32$ 并在表4分别报告他们的性能“K3”和“K32”。实验结果表明本文模型实现了在不同低维特征的维度设置下都相对稳定的性能，其中现在的维度设置 $K = 6$ 效果最好。

表 2所示的“互信息正则项”模块结构： 如章节3.2中讨论，“互信息正则项”模块由一个 3×3 卷积层和一个全连接层组成。也可以将其直接实现为显著性编码器的输出。特别地，我们可以输入 RGB 特征和深度特征到两个全连接层来分别获得 z_a 和 z_g。在表4中，本文报告了我们模型在该简单设置下的性能，标记为“SS”。我们观测到性能下降，表明了引入更多非线性来有效提取每个模态的特征表示的必要性。

互信息正则项的权重： 互信息正则项的权重 λ 控制互补信息的等级。本文中设置 $\lambda = 0.1$ 来实现平衡的训练。我们测试了不同的 λ 下的模型表现，并分别设置 $\lambda = 0$ 和 $\lambda = 1$。我们在表4中展示了这些版本的结果，用“W0”和“W1”标识。“W0”版本更差的性能表明了互补信息建模策略的有效性。此外，与本文在表2中的性能相比，我们观察到“W1”相对差的性能，这启发我们进一步探索以找到一个优化的互信息正则项权重。

5.4. 讨论

互信息最小化作为正则项的有效性： 本文分别对表 4中“W0”设置（没有互信息最小化作为正则项）和本文的方法在 NLPR 测试集上计算最高阶段特征嵌入（z_a^4 和 z_g^4）的平均绝对余弦相似度，分别是 $cosine(z_{a,g}^{M0}) = 0.90$ 和 $cosine(z_{a,g}^{Ours}) = 0.11$。这显著地表明了本文方案在对每个模态提取出更少相关特征上的优势。补充材料可见学得的特征嵌入可视化。

融合策略： 本文产生了四个不同的显著性图像作为中间输出，包括来自 RGB 分支（P_a）和深度分支（P_g）显著性预测，特征嵌入融合分支（P_f）的输出，和本文的通过融合 P_a，P_g 和 P_f 获得的最终预测 P。因为 P_f 已经包括了 z_a 和 z_g 的互补信息，本文定义 P_f 为不通过最终融合获取的最终预测结果 P。该性能展示在表 4中“P_f”栏。我们观察到相比本文的最终预测结果，“P_f”有更差的表现。主要原因是 z_a 和 z_g 是关于 RGB 图像和深度数据的高层特征嵌入，其主要捕获语义信息。将 z_a 和 z_g 直接融合会产生有更低结构性准确率的显著性预测结果。

深度数据的贡献： 显著性检测可以仅通过 RGB 图像实现。如在章节 1中所讨论的，深度数据为显著性检测引入了有用的几何信息。为了验证这个结论，我们在有深度数据和无深度数据的两种输入下训练我们的模型（只包括图 2中的编码器和解码器）。该性能展示在表4中“S_{rgbd}” 和 “S_{rgb}”栏。相比“S_{rgb}”，“S_{rgbd}”的更优性能表明了深度数据对于显著性检测的贡献。在补充材料中展示了深度数据如何贡献到显著性检测的例子。

深度生成： 本文根据 Holo50K 生成 COME15K 数据集，Holo50K [18] 中的立体对没有被严格修正，即使用目前最优的立体匹配算法 [6]，也可能导致严重的匹配失败。为了解决这个问题，本文在立体匹配算法中将水平搜索放宽至水平和垂直搜索，但只将水平变化作为立体差异。本文使用了一个修正后的立体匹配算法 [40] 来生成数据集中的差异/深度。此外，由于双目照相机被广泛用于移动设备，这使得它更容易对室内和室外场景获取深度信息。

带有深度数据的模型是早融合模型，其中深度数据和 RGB 图像在输入层被拼接。

表 5. 弱监督显著性检测基准模型的性能。

Method	NJU2K[21]				SSB[32]				NLPR [34]				SIP [11]				COME15K-Normal				COME15K-Difficult			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
Scribble	.823	.806	.869	.080	.820	.803	.884	.073	.820	.737	.863	.058	.815	.793	.888	.076	.802	.780	.856	.082	.767	.749	.812	.115
Polygon	.847	.827	.896	.065	.853	.831	.913	.056	.848	.789	.899	.043	.846	.822	.909	.060	.827	.805	.884	.065	.786	.774	.841	.096

表 6. 在新数据集 COME15K 中测试集上的性能。

	Metric	UCNet [46]	JLDCF [15]	A2dele [36]	DMRA [35]	CPFP [54]	S2MA [30]	CoNet [19]	BBS-Net [12]	CMINet Ours
Normal	$S_\alpha\uparrow$	.894	.894	.833	.782	.795	.877	.820	.902	**.915**
	$F_\beta\uparrow$	.883	.875	.835	.744	.716	.829	.796	.879	**.893**
	$E_\xi\uparrow$	.929	.919	.882	.812	.801	.881	.850	.923	**.941**
	$\mathcal{M}\downarrow$	.036	.042	.060	.105	.104	.059	.082	.039	**.033**
Difficult	$S_\alpha\uparrow$	.822	.845	.787	.743	.770	.828	.779	.853	**.867**
	$F_\beta\uparrow$	.814	.832	.795	.724	.704	.789	.774	.834	**.852**
	$E_\xi\uparrow$	.859	.870	.838	.775	.776	.836	.813	.876	**.893**
	$\mathcal{M}\downarrow$	.079	.075	.092	.137	.131	.092	.113	.071	**.064**

表 7. 立体显著性检测基准模型的性能。

NJU2K[21]			NJU400[20]			COME15K-Normal			COME15K-Difficult		
$S_\alpha\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
.874	.851	.056	.882	.851	.044	.874	.855	.047	.825	.812	.080

5.5. COME15K 上新的基准

本文提供了目前最优模型在新训练集上训练后新的基准，并且展示在表6中。此外，伴随有如图 3所示的丰富标注，本文还讨论了另外三个用于全监督/弱监督学习的基准。

基准 #1：在新训练集上重新训练现有的 RGB-D 显著性模型。本文划分测试集为一个中等难度的测试集（“普通”）和一个有难度的测试集（“困难”），分别有 4,600 和 3,000 个图像对。为了测试现有的 RGB-D 显著性检测模型在新测试集上的性能，我们用新训练集重新训练现有的 RGB-D 显著性检测模型，并在表6中展示了他们在新测试集上的性能。目前技术间的性能差距说明了本文数据集在模型学习和评价中的有效性。

基准 #2：立体显著性检测。因为本文的 RGB-D 显著性数据集基于一个立体数据集 [18] 构造，所以我们直接训练一个基于立体图像对的显著性目标检测模型，其中深度数据是隐式的而非显式从立体图像对中获得。虽然目前存在一些立体显著性检测模型 [23, 8, 32, 13]，但他们都将 RGB 图像和深度数据作为输入。与 [51] 相似，本文也设计了一个真实的立体显著性检测模型，以及提供了基准来表明本文数据集在立体显著性检测方面的潜力。我们在新的训练集上训练立体显著性检测模型，其中从左向右视图的图像作为输入，并且右视角下的真实显著性图像被用作监督标签。图 2中相同的编码器和解码器被用于我们的立体显著性检测模型。特别地，本文通过使用显著性编码器对左右视角图像间的代价立方体（cost volume）隐式地建模了几何信息。该性能被展示在表7中。我们在补充材料中解释了该架构和其他立体显著性数据集。

基准 #3 和 #4：笔画/多边形作为监督标签。对于笔画监督，本文遵循 [48]，并使用平滑损失和一个辅助的边缘检测分支作为约束，从而在预测中维持结构信息。本文通过在输入层中拼接 RGB 和深度数据训练笔画监督的 RGB-D 显著性检测模型，并且将拼接后的特征传入一个 3×3 的卷积层来调整模型 [48]。基于笔画标注的基准模型性能在表5中标识为“Scribble”。多边形标签通过多数派投票的方式产生。图 3（g）展示了多边形标签相比笔画方法覆盖了更大的有更好的结构信息的区域。本文通过采取图2中的模型，直接以多边形标注作为伪标签训练。并且在表5 “Polygon”栏提供了该基准模型的性能。

基准分析：表6中的 RGB-D 显著性基准展示了本文方法的优越性能。此外，目前最优方法间的差异说明了本文新测试数据集在模型评价中的有效性。表7中本文的立体显著性基准引入了另一个方法来隐式的使用几何信息。在表 5中本文的两个弱监督基准为弱监督 RGB-D 显著性检测任务提供了新的选择。

6. 结论

本文提出了一个基于多阶段级联学习的 RGB-D 显著性检测框架，它显式地建模 RGB 图像和深度数据之间的互补信息。通过在训练中最小化两个模态间的互信息，本文模型可以聚焦于每个模态的多样性的部分，而不是冗余信息。在这种方式下，本文的模型能够更有效地探索多模态信息。此外，本文引入了最大的 RGB-D 显著性检测数据集，有五种标注类型，以促进全/弱/无监督 RGB-D 显著性检测任务领域的发展。七个数据集上的四个新的基准和新数据集均证明了该模型相比于现有的 RGB-D 显著性检测技术的优越性。

本文定义了只接受从左到右视图图像作为输入的立体显著性模型为“真实的”立体显著性检测模型。

参考文献

[1] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3051–3060, 2018.

[2] Hao Chen and Youfu Li. Three-stream attention-aware network for RGB-D salient object detection. *IEEE T. Image Process.*, pages 2825–2835, 2019.

[3] 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 Recognit.*, 86:376–385, 2019.

[4] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE T. Image Process.*, 30:2313–2324, 2021.

[5] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for rgb-d salient object detection. In *Eur. Conf. Comput. Vis.*, 2020.

[6] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 22158–22169, 2020.

[7] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *ACM ICIMCS*, pages 23–27, 2014.

[8] 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 Processing Letters*, 23(6):819–823, 2016.

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Int. Conf. Comput. Vis.*, pages 4548–4557, 2017.

[10] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 2021.

[11] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *IEEE T. Neural Netw. Learn. Syst.*, 2020.

[12] 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 *Eur. Conf. Comput. Vis.*, 2020.

[13] Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin. Saliency detection for stereoscopic images. *IEEE T. Image Process.*, 23:1–6, 11 2013.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3146–3154, 2019.

[15] 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 Conf. Comput. Vis. Pattern Recog.*, 2020.

[16] 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 T. Cybern.*, pages 3171–3183, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[18] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2020.

[19] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *Eur. Conf. Comput. Vis.*, 2020.

[20] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE Int. Conf. Image Process.*, pages 1115–1119, 2014.

[21] Ran Ju, Yang Liu, Tongwei Ren, Ling Ge, and Gangshan Wu. Depth-aware salient object detection using anisotropic center-surround difference. *Signal Processing: Image Communication*, 38:115 – 126, 2015.

[22] Shuhui Wang Jun Wei and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI Conf. Art. Intell.*, 2020.

[23] Haksub Kim, Sanghoon Lee, and Alan Bovik. Saliency prediction on stereoscopic videos. *IEEE T. Image Process.*, 23:1476–90, 04 2014.

[24] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. Asif-net: Attention steered interweave fusion network for rgb-d salient object detection. *IEEE T. Cybern.*, 51(1):88–100, 2021.

[25] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. Rgb-d salient object detection with cross-modality modulation and selection. In *Eur. Conf. Comput. Vis.*, 2020.

[26] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for rgb-d salient object detection. In *Eur. Conf. Comput. Vis.*, 2020.

[27] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2806–2813, 2014.

[28] 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.

[29] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[30] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[31] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for rgb-d salient object detection. In *Eur. Conf. Comput. Vis.*, 2020.

[32] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 454–461, 2012.

[33] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *Eur. Conf. Comput. Vis.*, 2020.

[34] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: a benchmark and algorithms. In *Eur. Conf. Comput. Vis.*, pages 92–109, 2014.

[35] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Int. Conf. Comput. Vis.*, 2019.

[36] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[37] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

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

[39] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *Int. Conf. Learn. Represent.*, 2020.

[40] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong Li. Displacement-invariant matching cost learning for accurate optical flow estimation. In *Adv. Neural Inform. Process. Syst.*, 2020.

[41] Ningning Wang and Xiaojin Gong. Adaptive fusion for RGB-D salient object detection. *arXiv:1901.01369*, 2019.

[42] Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. In *Eur. Conf. Comput. Vis.*, 2018.

[43] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[44] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Int. Conf. Comput. Vis.*, 2019.

[45] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation

in street scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3684–3692, 2018.

[46] 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 Conf. Comput. Vis. Pattern Recog.*, 2020.

[47] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *International Conference on Computer Vision (ICCV)*, 2021.

[48] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[49] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In *Eur. Conf. Comput. Vis.*, 2020.

[50] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[51] Qiudan Zhang, Xu Wang, Shiqi Wang, Shikai Li, Sam Kwong, and Jianmin Jiang. Learning to explore intrinsic saliency for stereoscopic video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019.

[52] Zhao Zhang, Zheng Lin, Jun Xu, Wenda Jin, Shao-Ping Lu, and Deng-Ping Fan. Bilateral attention network for rgb-d salient object detection. *arXiv preprint arXiv:2004.14582*, 2020.

[53] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[54] 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 Conf. Comput. Vis. Pattern Recog.*, 2019.

[55] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *Eur. Conf. Comput. Vis.*, 2020.

[56] Yiran Zhong, Charles Loop, Wonmin Byeon, Stan Birchfield, Yuchao Dai, Kaihao Zhang, Alexey Kamenev, Thomas Breuel, Hongdong Li, and Jan Kautz. Displacement-invariant cost computation for efficient stereo matching. *arXiv preprint arXiv:2012.00899*, 2020.