UC-Net：不确定性启发的基于条件变分自动编码器的 RGB-D 显著性检测

张静 [1,4,5] 范登平 [2,6,*] 戴玉超 [3] Saeed Anwar[1,5]
Fatemeh Sadat Saleh[1,4] 张同 [1] Nick Barnes[1]
[1] 澳洲国立大学 [2] 南开大学计算机学院 [3] 西北工业大学 [4] ACRV [5] Data61 [6] IIAI

摘要

本文提出了首个通过数据标注过程进行学习，并将不确定性用于 *RGB-D* 显著性检测的框架。现有的 *RGB-D* 显著性检测方法将显著性检测任务视为点估计问题，并按照确定的学习流程生成单个显著性图。受显著性数据集标注过程的启发，本文提出了基于条件变分自动编码器的 *RGB-D* 显著性检测概率网络来模拟人类在图片标注时的不确定性，并通过在隐空间中采样的方式为每张输入图像生成多张显著性图。通过提出的“显著性一致”模型，本文可以将多张显著性图合并成一张准确的显著性图。在 *6* 个具有挑战的基准数据集上对 *18* 个有竞争力的算法进行定量和定性评估，证明本文方法在学习显著性图分布方面的有效性。代码和结果详见：*https://github.com/JingZhang617/UCNet* [1]。

1. 引言

目标级视觉显著性检测目的是将背景中最引人注意的显著性目标分离出来 [26, 2, 55, 63, 37, 28, 62]。最近，由于深度信息在人类视觉系统中的重要性以及深度传感技术的普及，通过 RGB-D 图像进行的视觉显著性检测引起人们广泛关注 [41, 64]。给定一对 RGB-D 图像，RGB-D 显著性检测的任务旨在通过探索彩色图像和深层数据之间的额外信息来预测并生成显著性图。

事实上，RGB-D 显著性检测是使用相应的基准数据集提供的显著性真值图来训练一个深度神经网络，其中显著性真值图来自人类共识或数据集创建者 [17]。在大规模 RGB-D 数据集的基础上，基于深度卷积神经网络的模型 [20, 41, 6, 24] 在学习从 RGB-D 图像对映射到其相应的显著性真值图方面取得了重大进展。考虑到当前 RGB-D 显著性检测的进展，本文认为该模式未能捕捉到显著性真值图标注时的不确定性。

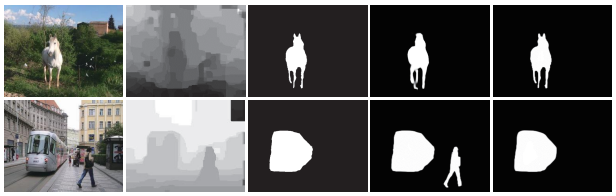

图 1. 提供了真值图与 *UC-Net* 预测的显著性图的比较。具有单个显著对象的图像（第一行），可以产生一致的预测结果。当存在多个显著物体时，本文方法可以得到不同的预测（第二行）。

人类视觉感知研究 [32] 表明，视觉显著性检测在某种程度上具有主观性。每个人在标注显著性图时可能都有其特定的偏好（之前在特定用户显著性检测 [25] 中已经讨论过）。现有的 RGB-D 显著性检测方法将显著性检测作为点估计问题，并按照既定的学习流程为每个输入图像生成单个显著性图，这种方式无法捕捉显著性的随机特征，且可能生成一个有偏的显著性模型，如图1第二行所示。本文提出的网络旨在研究可以产生多个预测（分布估计）的网络，从而反映出显著性的“主观性”的特性。受到人类感知的不确定性启发，本文提出了一个基于条件变分自动编码器 [50]（CVAE）的 RGB-D 显著性检测模型 *UC-Net*。它通过将输出空间的分布建模为一个生成模型，并以输入的 RGB-D 图像为条件来考虑标注中的人为不确定性，从而产生多个显著性预测。

*通讯作者：范登平 *(dengpfan@gmail.com)*
[1]本文为 CVPR2020 论文 [61] 的中文翻译版。

然而，在应用概率框架之前还存在一个问题，即现有的 RGB-D 基准数据集通常只为每对 RGB-D 图像提供一个显著性真值图。为了得到多样化和精确的预测[2]，本文遵循方向转移理论 [25]，采用了“躲藏和发现”原理 [49]，通过迭代将 RGB 图像中的显著前景隐藏并测试，迫使深层网络学习具有多样性的显著性图。通过迭代隐藏策略为每对输入的 RGB-D 图像生成多个显著性映射，反映了人类标注的多样性/不确定性。

此外，RGB-D 显著性数据集中的深度数据可能有噪声，并且 RGB 图像和深度信息的直接融合可能会让网络为适应噪声而不堪重负。针对深度噪声问题，本文提出了深度修正网络作为辅助元件来产生具有丰富语义和几何信息的深度图。本文还引入了显著性一致模块来模拟生成显著性真值的多数投票机制。

本文的主要贡献是：1）提出了 RGB-D 显著性预测条件概率模型，该模型可以产生不同的显著性预测而非单一的显著性图；2）通过显著性一致模块提供一种能更好地模拟显著性检测工作的机制；3）提出深度修正网络来减少深度数据中固有的噪声；4）在 6 个 RGB-D 显著性检测基准数据集上的大量实验结果证明了 *UC-Net* 的有效性。

2. 相关工作

2.1. RGB-D 显著性检测

根据 RGB 图像与深度图像之间互补信息的融合方式，现有的 RGB-D 显著性检测模型大致可分为三种：早期融合模型 [43]、后期融合模型 [54, 24] 和跨层融合模型 [41, 5, 7, 6, 64]。[43] 提出了早期融合模型，为每对 RGB-D 图像的每个超像素生成特征，然后将其输入至卷积神经网络以产生每个超像素的显著性特征。[54] 引入了一种新的融合网络（AFNet）来自适应地融合 RGB 图像和深度分支。[24] 以相似的流程通过全连接层融合 RGB 图像和深度信息。[7] 采用多尺度多路径网络融合不同模态的信息。[5] 提出了基于互补感知的 RGB-D 显著性检测模型，该模型用一个互补感知融合块来融合各模式同一阶段的特征。[6] 还提出了用于多模态融合的注意力感知跨层组合块。[64] 在增强深度线索之前整合了对比度，并采用流体金字塔整合框架实现了多尺度跨模态的特征融合。为了有效地将几何和语义信息集成到循环学习框架中，[41] 引入了深度诱导的多尺度 RGB-D 显著性检测网络。

[2]预测的多样性与图像的内容有关。具有简单内容的图像可以产生一致的预测（图1中第一行），而复杂的图像可能会产生不同的预测（图1中第二行）。

2.2. 基于 VAE 或 CVAE 的深度概率模型

自从 Kingma 等人 [30] 和 Rezende 等人 [45] 的开创性工作以来，变分自动编码器（VAE）及其条件性约束的 CVAE[50] 在各种计算机视觉问题中得到了广泛应用。为了训练 VAE，需要一个重建损失函数来衡量预测和真值的差异，以及一个正则化项来减少隐变量先验和后验分布间的分歧。VAE 一般定义隐变量服从特定的高斯分布，而 CVAE 将隐变量先验分布定义为条件分布，即其利用输入观测值调节隐变量的先验分布，从而产生输出。在低层视觉中，VAE 和 CVAE 已被应用于图像背景建模 [33]、锐化样本的隐空间表示 [21]、运动模式差异 [57]、医学图像分割模型 [3]、对图像固有的歧义进行建模 [31] 等任务。同时，VAE 和 CVAE 在不确定未来预测 [1, 53]、人体运动预测 [47] 和形状引导图生成 [11] 等更复杂的视觉任务中得到了广泛的应用。近年来，VAE 算法已经扩展到 3D 领域的目标定位应用中如三维网格变形 [52]、点云实例分割 [59] 等。

据我们所知，CAVE 还未被用于显著性检测。虽然 Li 等人 [33] 在显著性预测框架中采用了 VAE，但仅使用其对图像背景进行建模且通过重建残差区分显著目标与背景。而本文使用 CVAE 来模拟标签变量，表明人类标注的不确定性。本文是首次考虑到人类标注时的不确定性而在显著性预测网络中使用 CVAE 的。

3. 模型

本节介绍的是基于条件变分自动编码器的 RGB-D 显著性检测概率模型，该模型学习的是显著性图的分布而非单个预测。设 $\xi = \{X_i, Y_i\}_{i=1}^{N}$ 为训练数据集，其中 $X_i = \{I_i, D_i\}$ 表示 RGB-D 输入（由 RGB 图像 I_i 和深度图 D_i 组成），Y_i 表示真值显著性图。本文的模型在训练和测试期间的整个流程分别如图2和图3所示。

本文的网络主要分为 5 个模块：1）隐藏层网络（先验网络和后验网络）将 RGB-D 图像输入 X_i（用于先验网络）或者 X_i 和 Y_i（用于后验网络）映射到低维隐变量 $z_i \in \mathbb{R}^K (K$ 是隐空间的维数)；2）深度修正网络，它以 I_i 和 D_i 为输入，生成精确的深度图像 D_i'；3）显著性检测网络，它将 RGB 图像 I_i 和精确深度图像 D_i'

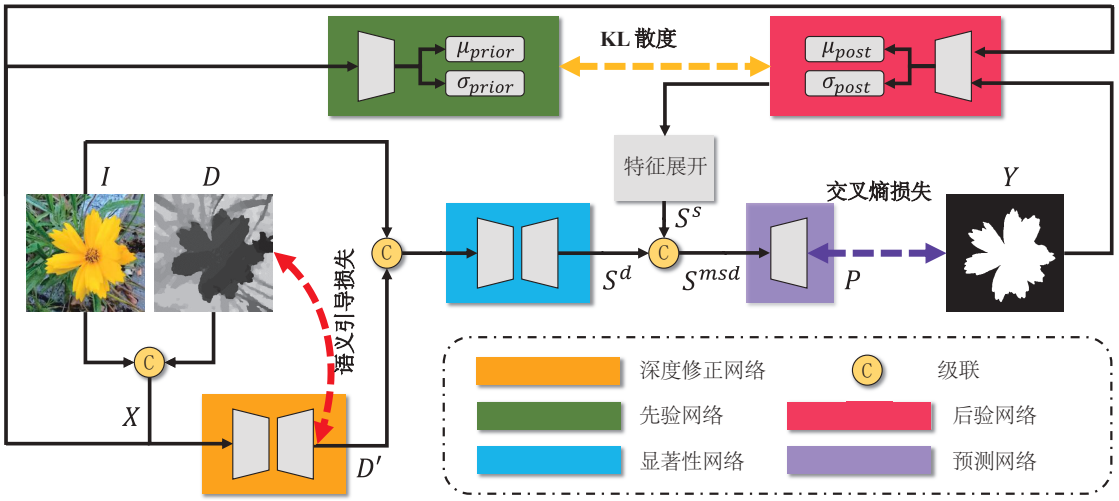

图 2. 网络训练流程。包括四个主要模块，即隐藏层网络（先验网络 $(\mu_{\text{prior}}, \sigma_{\text{prior}})$ 和后验网络 $(\mu_{\text{post}}, \sigma_{\text{post}})$），显著性网络，深度修正网络还有预测网络。隐藏层网络将 RGB-D 图像对 X（或与真值 Y 一起用于后验网络）映射到低维高斯隐变量 z。深度修正网络通过语义引导损失来细化原始深度图。该网络以 RGB 图像和细化后的深度图作为输入，生成显著性特征图。预测网络采用随机特征和确定性特征生成最终的显著性图。如图3所示，本文在测试阶段进行显著性一致步骤，该步骤根据真值显著性图生成机制来生成最终的显著性图。

图 3. 网络测试流程。本文对先验网络进行多次采样，以产生多样化和准确的预测。然后，显著性一致模块用于获得最终预测的多数投票结果。

映射到显著性特征图 S_i^d 中；4）预测网络，它利用隐藏层网络的随机特征 S_i^s 和显著性网络的确定特征 S_i^d 来生成显著性预测图 P_i；5）测试阶段的显著性一致模块用以模拟显著性真值的生成机制，使用单个提供的显著性真值图 Y_i 来评估性能。如下将介绍每个模块。

3.1. 基于 CVAE 的 RGB-D 显著性概率模型

条件变分自动编码器（CVAE）以输入数据 X 为条件参数将隐变量的先验分布调整为高斯分布。条件生成模型中一般包括三类变量：条件变量 X（设置的 RGB-D 图像对）、隐变量 z 和输出变量 Y。隐变量定义为 $P_\theta(z|X)$，联合输入条件变量 X，我们获得输出变量 Y 的分布 $P_\omega(Y|X,z)$，其中 z 的后验分布可表示为 $Q_\phi(z|X,Y)$。CVAE 的损失定义为：

$$\begin{aligned}\mathcal{L}_{\text{CVAE}} = E_{z\sim Q_\phi(z|X,Y)}[-\log P_\omega(Y|X,z)] \\ +D_{KL}(Q_\phi(z|X,Y)||P_\theta(z|X)),\end{aligned} \tag{1}$$

式中 $P_\omega(Y|X,z)$ 是在给定隐变量 z 和条件变量 X 时 $P(Y)$ 的可能性，KL 散度 $D_{KL}(Q_\phi(z|X,Y)||P_\theta(z|X))$ 可以作为正则化损失来减小先验分布 $P_\theta(z|X)$ 和后验分布 $Q_\phi(z|X,Y)$ 之间的差距。CVAE 旨在对编码误差 $D_{KL}(Q_\phi(z|X,Y)||P_\theta(z|X))$ 下的对数似然度 $P(Y)$ 进行建模。本文遵循 CVAE 的标准做法 [50]，设计了一个基于 CVAE 的 RGB-D 显著性检测网络，并如下描述模型的各个部分。

隐藏层网络： 本文将 $P_\theta(z|X)$ 定义为先验网络，它将输入的 RGB-D 图像对 X，映射到低维隐空间，其中 θ 是先验网络的参数集。在网络结构相同和提供真值显著性地图 Y 的情况下，本文将 $Q_\phi(z|X,Y)$ 定义为后验网络，其中 phi 是后验网络参数集。在隐藏层网络（先验网络和后验网络）中，本文使用五个卷积层将输入的 RGB-D 图像 X(或后验网络的 X 和 Y 的串联) 映射到隐变量 $z \sim \mathcal{N}(\mu, \text{diag}(\sigma^2))$ 中，其中 $\mu, \sigma \in \mathbb{R}^K$ 表示隐变量的均值和标准差，如图4所示。

本文把先验和后验网络的参数集分别定义为 $(\mu_{\text{prior}}, \sigma_{\text{prior}})$ 和 $(\mu_{\text{post}}, \sigma_{\text{post}})$。公式(1)中的 KL 散度用于衡量先验分布 $P_\theta(z|X)$ 和后验分布 $Q_\phi(z|X,Y)$ 间分布不匹配的程度，换句话讲，KL 散度用于代表用 $Q_\phi(z|X,Y)$ 表示 $P_\theta(z|X)$ 时丢失的信息数量。CVAE

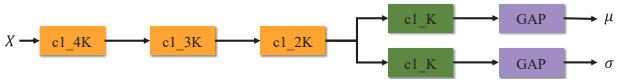

图 4. 隐藏层网络的详细结构，其中 K 为隐空间的维数，“c1_4K”表示卷积核为 1×1 输出通道大小为 $4K$ 的卷积层，“GAP”表示全局平均池化。

的典型应用涉及使用多个真值 Y [31] 来产生信息 $z \in \mathbb{R}^K$，z 中每个位置代表可能存在的标签变体或能致显著性标注多样的因素。由于只有一个真值，网络会单一匹配已有的标注 Y，所以直接使用提供的单一真值进行训练可能无法产生多样化的预测结果。

生成多种预测: 为了产生多样化且精确的预测，本文提出迭代隐藏技术 [49] 来生成更多标注，其灵感来于方向转移理论，如图5所示。本文利用训练集的均值以迭代的方式隐藏 RGB 图像中的显著区域。RGB 图像及其对应的真值被设置为“新标签生成”技术的开端。首先将显著物体的真值隐藏在 RGB 图像里，然后将修改后的图像反馈到现有的 RGB 显著性检测模型 [42] 中，生成显著性图并将其作为一个候选标注。本文对每个训练图像重复 3 次显著目标隐藏技术[3]以获得（包括真值在内）的 4 组不同的标注数据集，并且称这个数据集为“AugedGT”，即训练数据集。

训练期间，$Q_\phi(z|X,Y)$ 中的不同标注（如图5所示）可使先验网络 $P_\theta(z|X)$ 对给定输入 X 的标记变体进行编码。因为本文已通过隐藏技术得到了多样的标注，所以期待网络能够对复杂背景的图像产生多样的预测。测试过程中，每次采样都能获得通道数为 K 的一个随机特征 S^s(作为“预测网络”的输入) 如图3所示。

显著性网络: 本文设计了显著性网络，从输入的 RGB-D 数据中产生一个确定的显著特征映射 S^d，优化的深度数据来自深度修正网络。本文使用 VGG16[48] 作为编码器，并移除第五个池化层后的网络结构。为扩大感受野，本文采用 DenseASPP 算法 [58]，在 VGG16 网络的每阶上获得具有整个图像感受野的特征图。然后级联这些特征映射并传递到另一个卷积层以获得 S^d。显著网络的细节如图6所示，其中“c1_M”表示卷积核尺寸为 1×1 的卷积层，M 表示 S^d 的通道数。

特征扩展: 隐藏层网络（测试期间先验网络如图3中“采样”所示，后验网络如图2所示）的统计信息（尤其是 $(z \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2)))$）构成了特征扩展模块的输入。在

[3]本文发现，通常经过三次隐藏后，隐藏图像中不存在明显的物体。

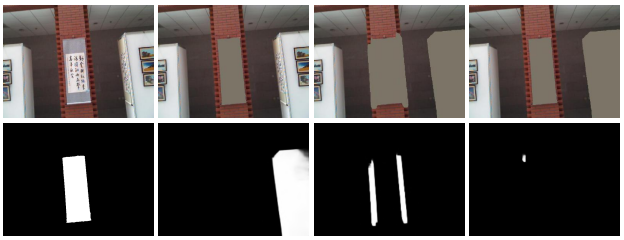

图 5. 新标签的生成。第一行：迭代隐藏的显著性预测区域，在第一幅图像中没有隐藏任何区域。第二行：隐藏图像的对应真值。

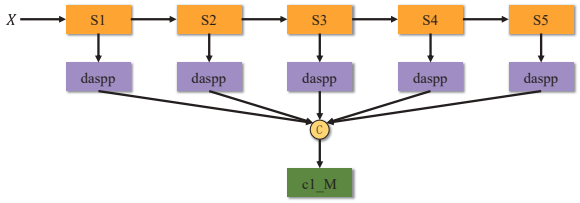

图 6. 显著性网络，其中“S1”代表了 VGG16 网络的第一阶段，“daspp”是 DenseASPP 的模块 [58]。

K 维向量的每个位置上给定一对 (μ^k, σ^k)，即可得到潜向量 $z^k = \sigma^k \odot \epsilon + \mu^k$，其中 $\epsilon \in \mathcal{N}(0, \mathbf{I})$。为与确定的显著性特征 S^d 融合，本文通过定义 ϵ 为二维高斯噪声图使 z^k 扩展为与 S^d 空间大小相同的特征图。当 $k = 1, ..., K$ 时，可得到 K（隐空间尺寸）通道的随机特征 S^s 来表示标注的不确定性。

预测网络: 隐藏层网络生成表示标签变量的随机特征 S^s，显著性网络以 X 为输入产生确定性的显著性特征 S^d。本文提出预测网络来融合上述分支如图2所示。S^s 和 S^d 的简单级联可能使网络只能从确定的特征中学习，无法对标记变量进行建模。受 [47] 的启发，本文将 S^s 和 S^d 按通道混合使网络无法区分确定分支和概率分支的特征。又将 S^d 和 S^s 级联，形成通道数为 $K + M$ 的特征图 S^{sd}。并且定义 $K + M$ 维变量 r（可学习参数）代表 $1, 2, ..., K + M$ 可能的排序，再根据 r 对 S^{sd} 进行智能通道混合得到混合特征 S^{msd}。预测网络用于将 S^{msd} 映射到单通道显著性图 P 中，预测网络包含 3 个输出通道数分别为 $K, K/2, 1$ 的 1×1 的卷积层。在测试过程中使用多个随机特征 S^s，就能以多次从隐藏层网络 $\mathcal{N}(\mu_{\text{prior}}, diag(\sigma^2_{\text{prior}}))$ 中对 S^s 进行采样的方式来获得多个预测。

3.2. 深度修正网络

RGB-D 显著性检测主要采用两种方法获取深度数据：一是通过微软 Kinect 等深度传感器，例如 DES [8]

和 NLPR[40] 数据集；二是从立体相机计算深度，例如 SSB[39] 和 NJU2K[27]。不管捕获技术如何，噪声都是深度数据中固有的。本文提出了一个语义引导的深度修正网络来产生如图2所示的精确深度信息，该网络被称为“深度修正网络”。深度修正网络的编码部分与“显著性网络”一致，解码部分则由四个连续的卷积层和双线性插值部分组成。假设深度图的边缘与 RGB 图像的边缘对齐，采用边界 IOU 损失 [38] 作为深度修正网络的正则化器，以 RGB 图像的强度为向导，来实现深度的精确化。那么深度修正网络的总损失定义为：

$$\mathcal{L}_{\text{Depth}} = \mathcal{L}_{sl} + \mathcal{L}_{\text{Ioub}}, \tag{2}$$

其中 $\mathcal{L}_{sl}$ 表示精确深度 D' 和原始深度 D 间的平滑损失 ℓ_1，$\mathcal{L}_{ioub}$ 是 RGB 图像 I 的精确深度 D' 和强度 Ig 间的边界 IOU 损失。给定预测的深度图 D' 和 RGB 图像的强度值 Ig，本文遵循 [38] 来计算 D' 和 Ig 的一阶导数，随后计算了 D' 和 Ig 梯度的幅值 gD' 和 gI，并将边界 IOU 损失定义如下：

$$\mathcal{L}_{\text{Ioub}} = 1 - 2\frac{|gD' \cap gI|}{|gD'| + |gI|}. \tag{3}$$

3.3. 显著性一致模块

显著性检测具有一定的主观性，通常需要多个注释者对一幅图像进行标注，通过多数投票策略得到最终的真值显著性图 [17]。尽管在显著性检测领域中，如何获得真值是众所周知的；但目前仍未有将该机制嵌入深度显著性框架的研究。现有模型将显著性检测定义为点估计问题而非分布估计问题。相反本文使用 CVAE 获取显著性分布。接下来将显著性一致方法嵌入概率框架，以计算测试阶段不同预测的多数投票结果，如图3所示。

测试过程中，本文用固定的 μ_{prior} 和 σ_{prior} 对先验网络进行抽样得到随机特征 S^s。用显著性网络中的每个 S^s 和确定性特征 S^d，得到显著性预测 P 的一个变体。为了得到 C 种不同的预测 $P^1, ..., P^C$，本文对先验网络进行 C 次采样。然后同时将多重预测输入至显著一致性模块中，来获得预测的一致性。

给定多个预测 $\{P^c\}_{c=1}^C$，其中 $P^c \in [0,1]$。我们先让 P^c 执行自适应阈值 [4] 来预测二进制[4]变量 P_b^c。每个像素 (u,v) 都得到 C 维特征向量 $P_{u,v} \in \{0,1\}$。本文定义 $P_b^{mjv} \in \{0,1\}$ 为代表 $P_{u,v}$ 多数投票的单通道显著性图。定义指标 $\mathbf{1}^c(u,v) = \mathbf{1}(P_b^c(u,v) = P_b^{mjv}(u,v))$ 来呈现二进制预测与多数预测是否一致。若 $P_b^c(u,v) = P_b^{mjv}(u,v)$，则 $\mathbf{1}^c(u,v) = 1$，否则 $\mathbf{1}^c(u,v) = 0$。经过显著性一致模块后可得灰度显著性图，如下所示：

$$P_g^{mjv}(u,v) = \frac{\sum_{c=1}^C \mathbf{1}^c(u,v)}{C} \sum_{c=1}^{C} (P_b^c(u,v)\} \times \mathbf{1}^c(u,v)). \tag{4}$$

3.4. 目标函数

在此阶段，损失函数由 $\mathcal{L}_{\text{CVAE}}$ 和 $\mathcal{L}_{\text{Depth}}$ 两部分组成。本文基于类间区分和类内相似性假设，提出以平滑损失为正则化器来实现边缘感知的显著性检测。继 [56] 之后，定义了显著性图在平滑项下的一阶导数为

$$\mathcal{L}_{\text{Smooth}} = \sum_{u,v} \sum_{d \in \vec{x}, \vec{y}} \Psi(|\partial_d P_{u,v}| e^{-\alpha |\partial_d Ig(u,v)|}), \tag{5}$$

其中 Ψ 被定义为 $\Psi(s) = \sqrt{s^2 + 1e^{-6}}$，$P_{u,v}$ 是 (u,v) 处的预测显著性图，$Ig(u,v)$ 是图像强度，d 表示在 $\vec{x}$ 和 $\vec{y}$ 上的偏导数。本文根据 [56] 设定 $\alpha = 10$。

平滑度损失（公式(5)）和边界 IOU 损失（公式(3)）都需要强度 Ig。根据 [60] 本文将 RGB 图像 I 转换为灰度强度图像 Ig:

$$Ig = 0.2126 \times I^{lr} + 0.7152 \times I^{lg} + 0.0722 \times I^{lb}, \tag{6}$$

其中 I^{lr}, I^{lg} 和 I^{lb} 表示从原始颜色空间移除伽马函数后线性颜色空间中的颜色分量。I^{lr} 通过以下方式实现：

$$I^{lr} = \begin{cases} \dfrac{I^r}{12.92}, & I^r \leq 0.04045, \\ \left(\dfrac{I^r + 0.055}{1.055}\right)^{2.4}, & I^r > 0.04045. \end{cases} \tag{7}$$

其中 I^r 是图像 I 的原始红色通道，本文用与公式(7)相同的方法计算 I^g 和 I^b。

本文的采用了平滑损失 $\mathcal{L}_{\text{Smooth}}$、深度损失 $\mathcal{L}_{\text{Depth}}$ 和条件变分自动编码损失 $\mathcal{L}_{\text{CVAE}}$，最终的损失函数定义如下：

$$\mathcal{L}_{\text{sal}} = \mathcal{L}_{\text{CVAE}} + \lambda_1 \mathcal{L}_{\text{Depth}} + \lambda_2 \mathcal{L}_{\text{Smooth}}. \tag{8}$$

本文实验中设定 $\lambda_1 = \lambda_2 = 0.3$。

训练细节： 本文将通道数 S^d 设为 $M = 32$，潜在空间的规模设为 $K = 8$。使用 Pytorch 对模型进行训

[4]作为真值图中的 $Y \in \{0,1\}$，本文生成一系列二值预测，每个预测值代表一种显著性定义的概率。

练，并在图片网络上进行 VGG16 参数预训练，初始化显著性网络和深度修正网络的编码器。新的权重层用 $\mathcal{N}(0, 0.01)$ 初始化，偏差设置为常数。并且使用了动量 0.9 的 Adam 方法，每一轮训练之后学习率降低 10%。基本学习率初始化为 1e-4。整个训练在一台装有英伟达精视 RTX 图形处理器的台式机上运行了 13 个小时，训练批量为 6 批，训练轮数为 30 次。输入尺寸为 352×352 的图像，测试阶段每幅图平均用时 0.06s。

4. 实验结果

4.1. 准备工作

数据集: 本文在六个数据集上进行了实验，包括五个广泛使用的 RGB-D 显著性检测测试数据集（名字分别为 NJU2K [27], NLPR [40], SSB [39], LFSD [34], DES [8]）和一个新发布的数据集 (SIP [17])。参考 [41] 训练数据集包括 NJU2K [27] 数据集的 1,485 张图像对和 NLPR [40] 数据集的 700 张图像。

模型比较: 将本文的方法与 18 种算法进行了比较，包括 10 种手工设计的传统模型和 8 种 RGB-D 显著性检测深度模型。

评价指标: 本文采用四个评价指标，两个广泛使用的指标：1) 平均绝对误差（MAE $\mathcal{M}$）;2) 平均 F 测度（F_β）以及两个最近提出的：3) 对比增强测度 (均值 E-measureE_ξ）[14]）4) 结构测度（S-measure, S_α）[13]。

4.2. 性能比较

定量比较: 表1中展示了本文方法和对比方法的性能。结果表明，本文的方法在所有的数据集上都取得了最佳性能，特别是在 SSB[39] 和 SIP[17] 上，本文的方法在 S-测度、E-测度和 F-测度上实现了显著的性能提升和 MAE 的大幅度下降。图7中给出了对比方法和本文的 E-测度和 F-测度曲线。可以观察到本文的方法不仅产生了稳定的 E-测度和 F-测度而且性能是最好的。

定性比较: 在图8中，本文展示了五幅图，图像展示的是本文方法的结果与一种新发布的 RGB-D 显著性检测方法（DMRA [41]）和两种广泛使用的产生结构化输出的方法进行比较的结果，这两种方法是 M-head[46] 和 MC-dropout[29]（在消融实验部分将详细讨论这两种方法）。通过用 M-head 和 MC-dropout 分别代替 CVAE，本文设计了基于 M-head 和 MC-dropout 的结构化显著性检测模型。图8的结果表明，本文的方法不仅可以对复杂背景下的图像进行高精度的预测（与 DMRA 相比 [41]），而且可以对复杂背景下的图像（图中第一行和最后一行所示）进行多种多样的预测（与基于 M-head[46] 和 MC-dropout[29] 的模型相比）。

4.3. 消融实验

本文进行了八个实验（如表2所示），来全面分析本文的框架，其中包括网络结构（“M1”, “M2”, “M3”），概率模型选择（“M4”, “M5”, “M6”），数据源选择（“M7”）以及新标签生成技术的有效性（“M8”））。当有结果比本文好时就将其加粗。

隐空间规模: 本文试图研究网络中高斯隐空间的维度 K 对网络的影响。经参数调整，发现 $K = 8$ 的效果最好。将 $K = 32$ 作为“M1”来展示性能。“M1”的性能比本文的结果差表明隐空间的规模是本文框架中的一个重要参数。本文进一步用 $K \in [2, 12]$ 进行了更多的实验，发现用 $K \in [6, 10]$ 的预测结果相对稳定。

深度修正网络的效果: 为了说明深度修正网络的有效性，故删除了这个分支，并将 RGB 图像和深度数据的级联反馈给显著性网络，如“M2”所示，它比本文的方法效果差。在 DES 数据集 [8] 上，本文提出的方法在 S-测度、E-测度和 F-测度上都取得了约 4% 的改进，这证明了深度修正网络的有效性。

显著性一致模块: 为了模拟显著性标记的过程，本文在测试过程中嵌入了显著性一致性模块（如图3所示），以获得多个预测的多数投票。将其从本文的框架中移除，并从先验网络 $P_\theta(z|X)$ 中随机抽取样本测试网络性能，以“M3”表示，与对比方法相校本文的效果是最好的。同时，在显著一致模块的嵌入下，本文取得的性能更好，这说明了显著性一致模块的有效性。

VAE 和 CVAE: 本文使用 CVAE 来模拟标记的不确定性，同时使用后验网络和先验网络来估计隐变量。为了检验 z 的先验分布为标准正态分布、后验分布为 $P_\theta(z|X)$ 时本文模型的表现，本文设计模型“M4”，发现其性能与 SOTA RGB-D 模型旗鼓相当。基于条件变分自动编码 [50] 模型进一步提升了“M4”的性能，证明了本文解决方案的有效性。

多头模型（M-head）和 CVAE: 多头模型 [46] 使用不同的解码器和一个共享的编码器产生多个预测，损失函数被定义为多个预测中最接近的一个。本文删除隐藏网络，并多次复制显著性网络的解码器来实现多

表 1. 在 6 个 RGBD 显著性数据集上，对 10 个领先的手工特征模型和 8 个深度模型进行了实验。↑ & ↓ 分别表示越大和越小越好，在此本文采用均值 F_β 和均值 E_ξ [14]。

		基于特征的手工制作模型										深度模型								*UC-Net*
	量度	LHM [40]	CDB [35]	DESM [8]	GP [44]	CDCP [66]	ACSD [27]	LBE [19]	DCMC [9]	MDSF [51]	SE [22]	DF [43]	AFNet [54]	CTMF [24]	MMCI [7]	PCF [5]	TANet [6]	CPFP [64]	DMRA [41]	Ours
NJU2K [27]	$S_\alpha \uparrow$	.514	.632	.665	.527	.669	.699	.695	.686	.748	.664	.763	.822	.849	.858	.877	.879	.878	.886	**.897**
	$F_\beta \uparrow$	.328	.498	.550	.357	.595	.512	.606	.556	.628	.583	.653	.827	.779	.793	.840	.841	.850	.873	**.886**
	$E_\xi \uparrow$	.447	.572	.590	.466	.706	.594	.655	.619	.677	.624	.700	.867	.846	.851	.895	.895	.910	.920	**.930**
	$\mathcal{M} \downarrow$	.205	.199	.283	.211	.180	.202	.153	.172	.157	.169	.140	.077	.085	.079	.059	.061	.053	.051	**.043**
SSB [39]	$S_\alpha \uparrow$	.562	.615	.642	.588	.713	.692	.660	.731	.728	.708	.757	.825	.848	.873	.875	.871	.879	.835	**.903**
	$F_\beta \uparrow$	.378	.489	.519	.405	.638	.478	.501	.590	.527	.611	.617	.806	.758	.813	.818	.828	.841	.837	**.884**
	$E_\xi \uparrow$	.484	.561	.579	.508	.751	.592	.601	.655	.614	.664	.692	.872	.841	.873	.887	.893	.911	.879	**.938**
	$\mathcal{M} \downarrow$	.172	.166	.295	.182	.149	.200	.250	.148	.176	.143	.141	.075	.086	.068	.064	.060	.051	.066	**.039**
DES [8]	$S_\alpha \uparrow$	.578	.645	.622	.636	.709	.728	.703	.707	.741	.741	.752	.770	.863	.848	.842	.858	.872	.900	**.934**
	$F_\beta \uparrow$	.345	.502	.483	.412	.585	.513	.576	.542	.523	.618	.604	.713	.756	.735	.765	.790	.824	.873	**.919**
	$E_\xi \uparrow$	.477	.572	.566	.503	.748	.613	.650	.631	.621	.706	.684	.809	.826	.825	.838	.863	.888	.933	**.967**
	$\mathcal{M} \downarrow$	.114	.100	.299	.168	.115	.169	.208	.111	.122	.090	.093	.068	.055	.065	.049	.046	.038	.030	**.019**
NLPR [40]	$S_\alpha \uparrow$	.630	.632	.572	.655	.727	.673	.762	.724	.805	.756	.806	.799	.860	.856	.874	.886	.888	.899	**.920**
	$F_\beta \uparrow$	.427	.421	.430	.451	.609	.429	.636	.542	.649	.624	.664	.755	.740	.737	.802	.819	.840	.865	**.891**
	$E_\xi \uparrow$	.560	.567	.542	.571	.782	.579	.719	.684	.745	.742	.757	.851	.840	.841	.887	.902	.918	.940	**.951**
	$\mathcal{M} \downarrow$	.108	.108	.312	.146	.112	.179	.081	.117	.095	.091	.079	.058	.056	.059	.044	.041	.036	.031	**.025**
LFSD [34]	$S_\alpha \uparrow$	.557	.520	.722	.640	.717	.734	.736	.753	.700	.698	.791	.738	.796	.787	.794	.801	.828	.847	**.864**
	$F_\beta \uparrow$	.396	.376	.612	.519	.680	.566	.612	.655	.521	.640	.679	.736	.756	.722	.761	.771	.811	.845	**.855**
	$E_\xi \uparrow$	.491	.465	.638	.584	.754	.625	.670	.682	.588	.653	.725	.796	.810	.775	.818	.821	.863	.893	**.901**
	$\mathcal{M} \downarrow$	.211	.218	.248	.183	.167	.188	.208	.155	.190	.167	.138	.134	.119	.132	.112	.111	.088	.075	**.066**
SIP [17]	$S_\alpha \uparrow$	.511	.557	.616	.588	.595	.732	.727	.683	.717	.628	.653	.720	.716	.833	.842	.835	.850	.806	**.875**
	$F_\beta \uparrow$	.287	.341	.496	.411	.482	.542	.572	.500	.568	.515	.465	.702	.608	.771	.814	.803	.821	.811	**.867**
	$E_\xi \uparrow$	.437	.455	.564	.511	.683	.614	.651	.598	.645	.592	.565	.793	.704	.845	.878	.870	.893	.844	**.914**
	$\mathcal{M} \downarrow$	.184	.192	.298	.173	.224	.172	.200	.186	.167	.164	.185	.118	.139	.086	.071	.075	.064	.085	**.051**

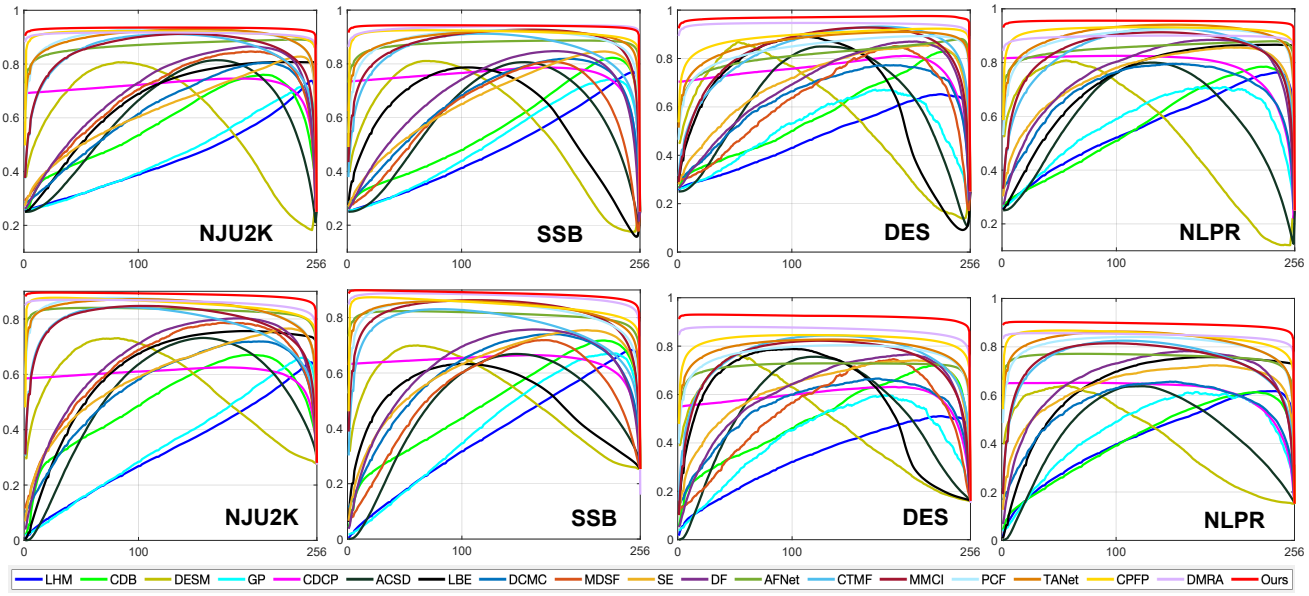

图 7. 在 4 个测试数据集上的 E-measure (1^{st} 行) 和 F-measure (2^{nd} 行) 曲线。

个预测（在本文中用“M5”表示）。本文将“M5”的表现作为多次预测的均值。“M5”优于 SOTA 模型（*e.g.*, DMRA），而基于多头的方法（“M5”）与基于 CVAE 的模型（*UC-Net*）仍存在差距。

MC-dropout 和 CVAE 的比较: MC-dropout[29] 在测试阶段使用随机淘汰的方式将随机性引入网络。本文遵循 [29]，并去掉隐变量的先验和后验网络，测试阶段在显著性网络的编码和解码部分使用随机 dropout。本文重复了 5 次该方法（随机淘汰率 = 0.1），并且生成的平均成绩为“M6”。与“M5”相似，“M6”的性能也优于 SOTA 模型，而基于 CVAE 模型的性能更为优越。

三通道法和深度图: 三通道法 [23] 是一种广泛使用的

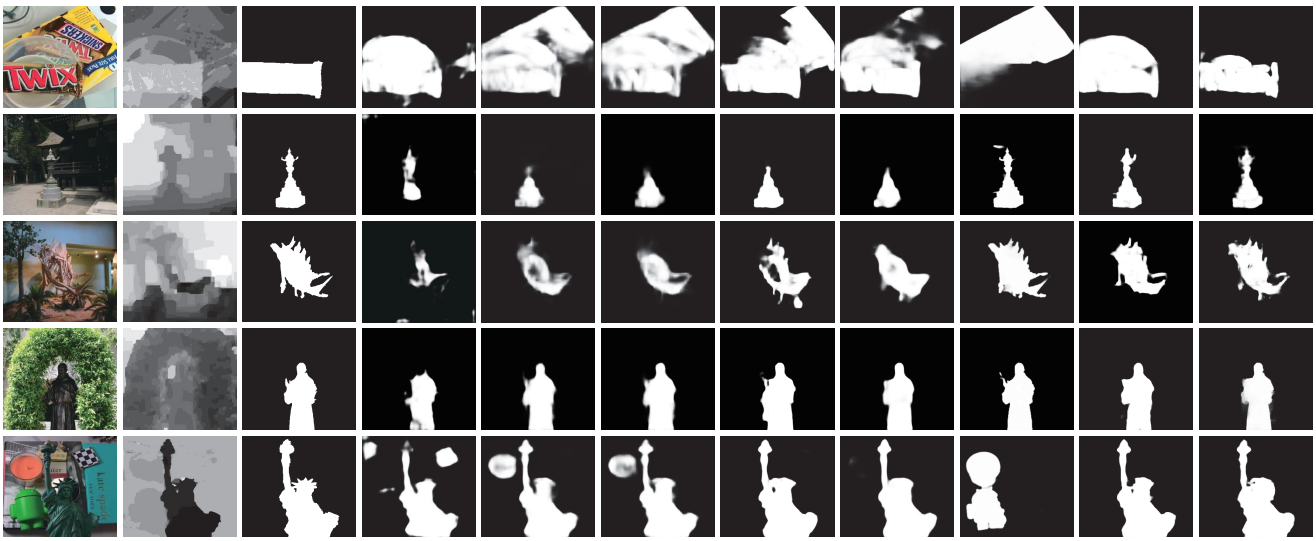

图 8. 显著性图的比较。“MH1”和“MH2”是多头模型的两个预测结果。“DP1”和“DP2”是测试期间两次 MC-dropout 的预测。“本文 (1)”和“本文 (2)”是本文基于条件变分自动编码模型的两个预测。对于具有分歧的图像，多头模型和 MC-dropout 模型基本上只能产生一致的输出，(如第 5^{th} 行所示)，而本文的模型可以产生多样化的预测结果。

表 2. RGB-D 显著性数据集的消融研实验。

	量度	*UC-Net*	M1	M2	M3	M4	M5	M6	M7	M8	M9
NJU2K [27]	$S_\alpha \uparrow$	.897	.866	.893	**.905**	.871	.885	.881	.893	.838	.866
	$F_\beta \uparrow$	.886	.858	**.887**	.884	.851	.878	.878	.884	.787	.812
	$E_\xi \uparrow$	.930	.905	.930	.927	.910	.923	.927	**.932**	.840	.866
	$\mathcal{M} \downarrow$	**.043**	.060	.046	.045	.059	.047	.046	.044	.084	.075
SSB [39]	$S_\alpha \uparrow$	**.903**	.854	.893	.900	.867	.891	.893	.898	.855	.872
	$F_\beta \uparrow$	**.884**	.831	.876	.868	.834	.864	.876	.882	.793	.805
	$E_\xi \uparrow$	**.938**	.894	.911	.922	.907	.921	.931	.934	.854	.870
	$\mathcal{M} \downarrow$	**.039**	.060	.043	.047	.057	.047	.043	.040	.073	.068
DES [8]	$S_\alpha \uparrow$	**.934**	.876	.896	.928	.897	.911	.896	.918	.811	.911
	$F_\beta \uparrow$	**.919**	.844	.868	.902	.867	.897	.868	.904	.724	.843
	$E_\xi \uparrow$	**.967**	.906	.928	.947	.930	.945	.928	.953	.794	.910
	$\mathcal{M} \downarrow$	**.019**	.035	.026	.024	.033	.024	.026	.023	.065	.036
NLPR [40]	$S_\alpha \uparrow$	**.920**	.878	.919	.918	.890	.899	.910	.915	.850	.883
	$F_\beta \uparrow$	.891	.846	**.897**	.878	.845	.875	.867	.889	.759	.795
	$E_\xi \uparrow$	.951	.911	**.953**	.941	.924	.937	.933	.951	.841	.883
	$\mathcal{M} \downarrow$	.025	.039	**.024**	.029	.037	.029	.028	.025	.057	.045
LFSD [34]	$S_\alpha \uparrow$	**.864**	.799	.847	.862	.820	.838	.847	.853	.729	.823
	$F_\beta \uparrow$	**.855**	.791	.838	.841	.802	.833	.838	.848	.661	.779
	$E_\xi \uparrow$	**.901**	.829	.879	.885	.865	.875	.879	.891	.720	.818
	$\mathcal{M} \downarrow$	**.066**	.101	.079	.075	.093	.079	.079	.073	.145	.108
SIP [17]	$S_\alpha \uparrow$	**.875**	.846	.867	.870	.851	.859	.867	.865	.810	.845
	$F_\beta \uparrow$	**.867**	.837	.860	.848	.821	.853	.860	.855	.751	.795
	$E_\xi \uparrow$	**.914**	.884	.908	.901	.893	.905	.908	.908	.816	.852
	$\mathcal{M} \downarrow$	**.051**	.068	.056	.059	.067	.057	.056	.056	.094	.079

技术，它将深度数据编码为三个通道：水平视差、离地高度及局部像素曲面法线和理论重力方向的夹角。为了获得更好的特征识别结果，三通道法被广泛应用于 RGB-D 相关的密集预测模型 [10, 24] 中。为了测试三通道法是否适用于本文的情况，用其替换深度图，性能如“M7”所示。本文发现三通道法代替原始深度数据取得了类似的性能。

AugGT 有效性: 为了产生不同的预测，本文遵循 [49] 为训练数据集生成不同的标签。为了说明该策略的有效性，本文以 RGB-D 图像作为输入，仅用显著性网络进行训练生成单通道显著图。“M8”“M9”分别表示使用训练数据集和增强的训练数据集。相比“M8”，“M9”性能上的优势证明了新的标签生成技术的有效性。

5. 总结

本文受人类标注时的不确定性启发，提出了首个名为 ***UC-Net*** 的不确定性网络，该网络基于条件变分自动编码器用于 RGB-D 显著性检测。现有的方法将显著性检测作为一个点估计问题，而本文提出学习显著性图的分布。本文的模型能够通过显著性一致模块生成多个在真值标注生成过程中被舍弃的标签。通过在六个标准且具有挑战性的数据集上进行定量和定性评估，其结果证明了本文在学习显著性图的分布方面的优势。在将来，我们希望将本文的方法扩展到其他显著性检测问题（例如, VSOD [18]、RGB SOD [12, 65]、Co-SOD [16]）。此外，我们计划获取具有多个人类注释的新数据集，以进一步建模交互式图像分割 [36]、伪装物体检测 [15] 等人类不确定性的统计。

参考文献

[1] Abubakar Abid and James Y. Zou. Contrastive Variational Autoencoder Enhances Salient Features. *CoRR*, abs/1902.04601, 2019.

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

[3] Christian F. Baumgartner, Kerem Can Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. In *MICCAI*, pages 119–127, 2019.

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient Object Detection: A Benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.

[5] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D Salient Object Detection. In *IEEE CVPR*, pages 3051–3060, 2018.

[6] Hao Chen and Youfu Li. Three-stream Attention-aware Network for RGB-D Salient Object Detection. *IEEE TIP*, pages 2825–2835, 2019.

[7] 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. *PR*, 86:376–385, 2019.

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

[9] 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 SPL*, 23(6):819–823, 2016.

[10] Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu. Translate-to-Recognize Networks for RGB-D Scene Recognition. In *IEEE CVPR*, pages 11836–11845, 2019.

[11] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A Variational U-Net for Conditional Appearance and Shape Generation. In *IEEE CVPR*, pages 8857–8865, 2018.

[12] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 186–202, 2018.

[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, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged Object Detection. In *IEEE CVPR*, 2020.

[16] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a Deeper Look at the Co-salient Object Detection. In *IEEE CVPR*, 2020.

[17] 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 TNNLS*, 2020.

[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] 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.

[20] Keren Fu 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*, 2020.

[21] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taïga, Francesco Visin, David Vázquez, and Aaron C. Courville. PixelVAE: A Latent Variable Model for Natural Images. In *ICLR*, 2016.

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

[23] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-

D images for object detection and segmentation. In *ECCV*, pages 345–360, 2014.

[24] 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 TCYB*, pages 3171–3183, 2018.

[25] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *VR*, 40(10):1489 – 1506, 2000.

[26] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998.

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

[28] Shuhui Wang Jun Wei and Qingming Huang. F3Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI*, 2020.

[29] Alex Kendall, Vijay Badrinarayanan, , and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In *BMVC*, 2017.

[30] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2013.

[31] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. In *NeurIPS*, pages 6965–6975, 2018.

[32] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, 2013.

[33] Bo Li, Zhengxing Sun, and Yuqi Guo. SuperVAE: Superpixelwise Variational Autoencoder for Salient Object Detection. In *AAAI*, pages 8569–8576, 2019.

[34] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *IEEE CVPR*, pages 2806–2813, 2014.

[35] 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.

[36] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive Image Segmentation with First Click Attention. In *IEEE CVPR*, 2020.

[37] Yi Liu, Qiang Zhang, Dingwen Zhang, and Jungong Han. Employing Deep Part-Object Relationships for Salient Object Detection. In *IEEE ICCV*, 2019.

[38] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-Local Deep Features for Salient Object Detection. In *IEEE CVPR*, 2017.

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

[40] 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.

[41] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced Multi-scale Recurrent Attention Network for Saliency Detection. In *IEEE ICCV*, 2019.

[42] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-Aware Salient Object Detection. In *IEEE CVPR*, 2019.

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

[44] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting Global Priors for RGB-D Saliency Detection. In *IEEE CVPRW*, pages 25–32, 2015.

[45] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *ICML*, pages 1278–1286, 2014.

[46] Christian Rupprecht, Iro Laina, Maximilian Baust, Federico Tombari, Gregory D. Hager, and Nassir Navab. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. In *IEEE ICCV*, pages 3611–3620, 2017.

[47] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, Stephen

Gould, and Amirhossein Habibian. Learning Variations in Human Motion via Mix-and-Match Perturbation. *arXiv e-prints*, page arXiv:1908.00733, 2019.

[48] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2014.

[49] Krishna Kumar Singh and Yong Jae Lee. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. In *IEEE ICCV*, 2017.

[50] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In *NeurIPS*, pages 3483–3491, 2015.

[51] 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 TIP*, 26(9):4204–4216, 2017.

[52] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational Autoencoders for Deforming 3D Mesh Models. In *IEEE CVPR*, 2018.

[53] Jacob Walker, Carl Doersch, Harikrishna Mulam, and Martial Hebert. An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders. In *ECCVW*, pages 835–851, 2016.

[54] Ningning Wang and Xiaojin Gong. Adaptive Fusion for RGB-D Salient Object Detection. *IEEE Access*, 7:55277–55284, 2019.

[55] 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 CVPR*, 2019.

[56] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion Aware Unsupervised Learning of Optical Flow. In *IEEE CVPR*, 2018.

[57] Yan, Xinchen, Rastogi, Akash, Villegas, Ruben, Sunkavalli, Kalyan, Shechtman, Eli, Hadap, Sunil, Yumer, Ersin, Lee, and Honglak. MT-VAE: Learning Motion Transformations to Generate Multimodal Human Dynamics. In *ECCV*, pages 276–293, 2018.

[58] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. DenseASPP for Semantic Segmentation in Street Scenes. In *IEEE CVPR*, pages 3684–3692, 2018.

[59] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In *IEEE CVPR*, 2019.

[60] Shivanthan A. C. Yohanandan, Adrian G. Dyer, Dacheng Tao, and Andy Song. Saliency Preservation in Low-Resolution Grayscale Images. In *ECCV*, 2018.

[61] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8582–8591, 2020.

[62] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-Supervised Salient Object Detection via Scribble Annotations. In *IEEE CVPR*, 2020.

[63] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective. In *IEEE CVPR*, pages 9029–9038, 2018.

[64] 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*, 2019.

[65] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge guidance network for salient object detection. In *IEEE ICCV*, pages 8779–8788, 2019.

[66] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *IEEE ICCVW*, 2017.