PiCANet：针对显着性检测学习基于像素的上下文注意

年刘[1] 韩俊伟[1*] 杨明轩[2,3]
[1]西北理工大学 [2]加州大学默塞德分校 [3]Google Cloud
{liunian228, junweihan2010}@gmail.com mhyang@ucmerced.edu

摘要

上下文在显着性检测任务中发挥重要作用。 但是，考虑到上下文区域，并非所有上下文信息都有助于最终任务。 在本文中，我们提出了一种新颖的基于像素的上下文关注网络，即PiCANet，以学习如何选择性地关注每个像素的信息上下文位置。 具体而言，对于每个像素，其可以生成每个注意权重对应于每个上下文位置处的上下文相关性的注意力映射。 然后可以通过有选择地聚合上下文信息来构建出席的上下文特征。我们分别以全球和当地的形式制定提议的PiCANet，分别参与全球和当地情况。 这两种模式都是完全可以区分的，可以嵌入CNN进行联合培训。 我们还将所提出的模型与U-Net架构结合起来，以检测显着对象。 大量的实验表明，提出的PiCANets可以持续改善显着性检测性能。 全球和当地的PiCANets分别有助于学习全球对比度和均匀度。 因此，我们的显着性模型可以更加准确和均匀地检测显着物体，从而更好地对付最先进的方法。

1. 介绍

显着性检测旨在建模人类视觉注意机制，以检测不同的区域或对象，人们可能将视线聚焦在视觉场景中。情境信息在这个视觉任务中起着至关重要的作用。 作为最早的先驱计算显着性模型之一，Itti等人 [12]计算每个像素与其周围区域之间的特征差异作为对比度来推断显着性。 随后开发了许多方法[7,4,15]，这些方法利用局部或全局情境作为评估每个图像位置的对比度（即局部对比度或全局对比度）的参考。 这些模型将所涉及的上下文区域的所有位置处的视觉信息聚合成上下文特征以推断对比度。

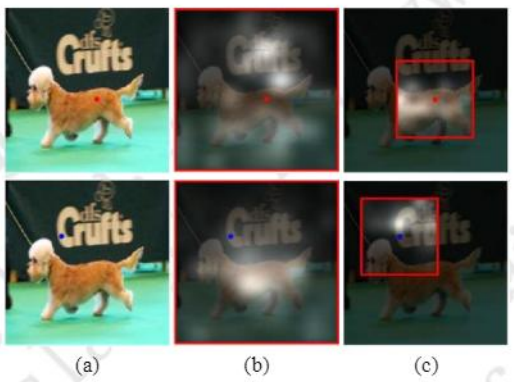


图1.学习的全局和局部像素相关注意映射示例。 （a）显示原始图像和两个示例像素，即前景狗上的红点和背景上的蓝点。 （b）和（c）分别显示了两个像素的学习到的全局和局部情景注意图。 每个位置的亮度表示其关注权重的大小。红色框表示所指的上下文区域。

最近，卷积神经网络（CNN）被引入到显着性检测中以学习有效的情境表示。 具体来说，几种方法[18,24,47]首先直接使用CNN从不同上下文的多个图像区域中提取特征，然后结合这些上下文特征来推断显着性。 其他一些模型[17,19,23,22,37,9,26,45,46,38]采用完全卷积网络（FCNs）[25]在每个图像位置进行特征表示，并以卷积方式生成显着图。 在这些模型中，第一个学校从每个输入图像区域提取上下文特征，而第二个学校从其相应的接受区域提取每个图像位置处的特征。

然而，所有现有的模型都利用上下文区域来构建上下文特征，其中每个上下文位置处的信息被整合。 直观地说，对于特定的图像像素，并非所有的上下文信息都有助于其最终决定。 一些相关区域通常更有用，而其他嘈杂的响应应该被丢弃。 例如，对于图1中第一行的红点像素，我们需要将其与背景进行比较以推断其全局对比度。 如果我们想要

*通讯作者

为了检查它是否属于前景狗，以便统一突出整条狗，我们需要参考狗的其他部分。 而对于第二行中的蓝点像素，我们需要分别参考前景狗和背景的其他部分。因此，如果我们能够识别相关的上下文区域并为每个像素构建信息化的上下文特征，则可以做出更好的决策。 尽管如此，现有方法尚未解决这个重要问题。

为了解决上面讨论的问题，在本文中，我们提出了一种新的Pixel-wise上下文注意网络，被称为PiCANet，以便为每个图像像素学习这些信息性上下文区域。 它通过为每个像素产生上下文关注来显着改善软注意模型[1]，这对于整个神经网络社区来说是一个真正的新颖想法。具体来说，如图1所示，所提出的PiCANet学习在每个像素的上下文区域上产生软注意力，其中注意力权重指示每个上下文位置与被引用像素的相关程度。 然后对来自上下文区域的特征进行加权并聚合以获得有人值守的上下文特征，其只考虑信息上下文位置而忽略每个像素的有害特征。 因此，所提出的PiCANets可以显着地促进显着性检测任务。

为了将上下文与不同的范围结合起来，我们分别以两种形式制定PiCANet：全局和本地PiCANet，分别有选择地整合全局上下文和本地上下文。 此外，我们的Pi-CANets实现完全可以区分。 因此，他们可以灵活地嵌入ConvNets并启用联合培训。

我们将全局和本地PiCANets分层嵌入到U-Net架构[30]中，该架构是一个具有跳过连接的编码器 - 解码器卷积网络，用于检测显着对象。 在解码器中，我们逐渐在多尺度特征地图上使用多个全局和本地PiCANet。 因此，我们从全局视图到局部上下文（从粗尺度到精尺度）构建出席的上下文特征，并使用它们来增强卷积特征以促进每个像素处的显着性推断。 图1显示了一些学习到的关注地图的例子。 对于每个像素（红色和蓝色的点），图1（b）中所示的全球学习注意力可以关注前景对象的背景和反之，这与全局对比机制完全匹配。 虽然图1（c）中显示的学习局部关注可以关注与其局部背景中与参照像素具有相似外观的区域，以使显着图更加均匀。

我们的贡献可以总结如下：

1. 我们提出新颖的PiCANet来产生对每个像素的上下文区域的关注。 因此，可以获得信息性的背景特征以促进最终决定。 此外，我们在两个方面都制定了PiCANet全球和当地的形式，分别参与全球和当地的情况，并且具有完全的可区分性，以便能够与ConvNets进行联合培训。

2. 我们通过将PiCANets嵌入到U-Net架构中提出了一种新颖的显着性检测模型。 PiCANets用于分层整合有人参与的全局上下文和多尺度本地上下文，这可以有效地提高显着性检测性能。

3. 在六个基准数据集上的广泛实验结果证明了与其他最先进的模型相比，所提出的Pi-Canets和显着性模型的有效性。 我们还深入分析并解释了为什么提议的PiCANets表现良好。

2. 相关工作

注意网络。 最近，神经网络引入了注意模型来模拟视觉场景中关注信息区域的视觉注意机制。 Mnih等人[28]提出了一个经常性关注模型和硬对齐。 但是，培养如此强硬的注意力模型是困难的。 随后，Bahdanau等。 [1]为机器翻译开发了具有可微软对齐的注意模型。 近年来，注意模型已应用于多项视觉任务。 Xu等人
[41]使用图像标题的经常性关注模型将单词与图像区域对齐。 在[32]中，Sermanet et al。 通过关注区分性区域采用经常关注的细粒度分类模型。 另外，引入注意模型用于回答视觉问题以关注与问题相关的图像区域[40,44]。 Li等人 [20]利用注意力去关注全球环境来指导物体检测。 这些作品表明，注意模型可以通过关注信息环境显着帮助计算机视觉任务。 然而，现有方法一次只考虑生成一个全局上下文关注映射，我们称之为图像上下文关注。 这些模型限制了卷积网络中关注网络的应用，特别是对于像素任务，因为不同的像素具有不同的信息情境区域。 在[3]中，Chen等人 为每个像素生成用于语义分割的注意力权重。然而，这种方法使用注意力来选择每个像素的多尺度特征上的自适应尺度，我们称之为像素尺度关注。 相反，我们提出的PiCANet会为每个像素的上下文区域产生注意力。

显着性检测。 传统的显着性模型主要依靠各种显着性提示来检测显着对象，包括局部对比[15]，全局对比[4]和背景优先[43]。 最近，随着CNN的使用，许多工作在显着性检测方面取得了可喜的成果。 接下来，我们简要回顾这些模型。

刘等人。 [24] Li和Yu [18]采用CNN来提取多尺度图像区域上的多尺度上下文特征，以分别推断每个像素和每个超像素的显着性。 同样，赵等人。 [47]在全球和当地情况下使用CNN。 在[19]中，基于FCN的显着性模型和基于多尺度图像区域的显着性模型相结合。 Wang等人 [37]反复采用FCN逐步改进显着图。 Liu和Han [23]使用基于U-Net的网络来分层预测并优化从全局视图到精细局部视图的显着图。 同样，罗等人。 [26]和张等人。 [45]也使用基于U-Net的模型来结合多级上下文来检测显着对象。 Wang等人 [38]还通过组合本地和全局上下文信息来逐步细化显着图。 在[9]中，短连接被引入到HED网络内的多尺度侧输出[39]，以提高显着性检测性能。 胡等人。 [10]建议采用基于水平集的损失来训练其显着性检测网络，并使用引导式超像素过滤来优化显着性图。

尽管现有的基于DNN的模型包含了用于显着性检测的各种上下文，但这些方法全部使用上下文区域。 通常情况下，工作在

[23,26,45]，它们具有类似的U-Net架构

我们在本文中使用的一种方法，通过不同的网络架构合并多尺度上下文，

最后整合来自其接受领域的信息。 相比之下，我们使用提议的PiCANets仅选择性地关注信息的上下文位置。 在[17]中，作者使用经常性关注模型来选择局部区域来优化其显着图。 然而，他们采用空间变换关注网络[13]在每个时间步选择一个细化区域，其模型仍然属于图像式关注类别。 相反，我们的PiCANets可以为每个像素生成柔和的上下文关注。

3. 像素智能上下文注意网络

所提出的PiCANet旨在在其上下文区域上的每个像素处生成注意映射，并构建出席的上下文特征以增强Convnets的特征可表示性。 给定一个卷积（Conv）特征映射FR$^{W\times H\times C}$，其中W，H，C分别表示其宽度，高度和通道数量，∈我们提出了两种基于像素的注意模式：全局注意和局部注意。 对于F中的每个位置（w，h），前者在整个特征图F上产生注意力，而后者在以（w，h）为中心的局部区域上工作。

3.1. 全球PiCANet

为了全球的关注，我们展示了网络架构 -
如图2（a）所示。 由于我们倾向于在每个像素的全局范围内产生注意力，因此我们需要使每个像素能够首先“看到”整个特征图F. 至为此，可以使用各种网络架构，其接受场是整个图像，例如完全连接的层。 在这里，我们采用更有效和更高效的ReNet模型[35]，该模型使用四个循环神经网络沿着两个方向水平和垂直扫描图像，以结合全球背景。 具体来说，如图2（a）中的橙色虚线框所示，双向LSTM（biLSTM）[6]首先沿着F的每一行部署，然后将每个像素的两个隐藏状态连接起来，使得每个像素都记住它的左右上下文。 接下来，沿获得的特征映射的每一列部署另一个biLSTM，使得每个像素能够记住它的顶部和底部上下文。 通过交替地水平和垂直扫描，来自四个方向的上下文可以被混合，这将每个像素的信息传播到所有其他像素。 因此，全局上下文被有效地结合在每个像素处。

接下来，我们使用香草Conv层将ReNet特征映射转换为D声道，其中D = W H。然后，在每个像素（w，h）处，所获得的特征向量表示为x^{w,h}通过softmax函数进行标准化以生成全局关注权重α^{w,h}：

$$\alpha^{w,h}_{i} = \frac{\exp(x^{w,h}_{i})}{\sum_{j=1}^{D}\exp(x^{w,h}_{j})}, \quad (1)$$

上下文位置（W$_i$，H$_i$）处的上下文相关性，其中i 1，D，x^{w,h}，α^{w,h}R和α^{w,h}）与所涉及的像素（w，h）相关。

最后，如图2（b）所示，对于像素（w，h），F中所有位置的特征用α^{w,h}进行加权求和以构造出席的上下文特征F$_{全}$：

$$F^{w,h}_{全} = \sum_{i=1}^{D}\alpha^{w,h}_{i} f_i \quad (2)$$

其中f$_i$ R^{C}是F和W中（W$_i$，H$_i$）的Conv特征F$_{全}$与F具有相同的尺寸。

3.2. 本地PiCANet

至于本地注意力，在每个像素（w，h）处，我们只对以（w，h）为中心的局部邻域上下文区域进行参与操作，形成局部特征立方体F^{W,H}R$^{W\times H\times C}$，宽度W 和高度H 。 网络架构如图2（c）所示。 再次，我们首先需要每个像素来“看”W H 上下文区域。 我们只是使用Conv层来达到这个目的。 具体来说，我们在F上部署了几个Conv层
它们的感受野达到W ×H 的大小。 然后，如与全球PiCANet相同，Conv层被用于trans-
形成D = W H 通道的结果特征映射。 接下来，通过softmax标准化（类似于（1））也生成本地关注权重α^{w,h}。 最后，如

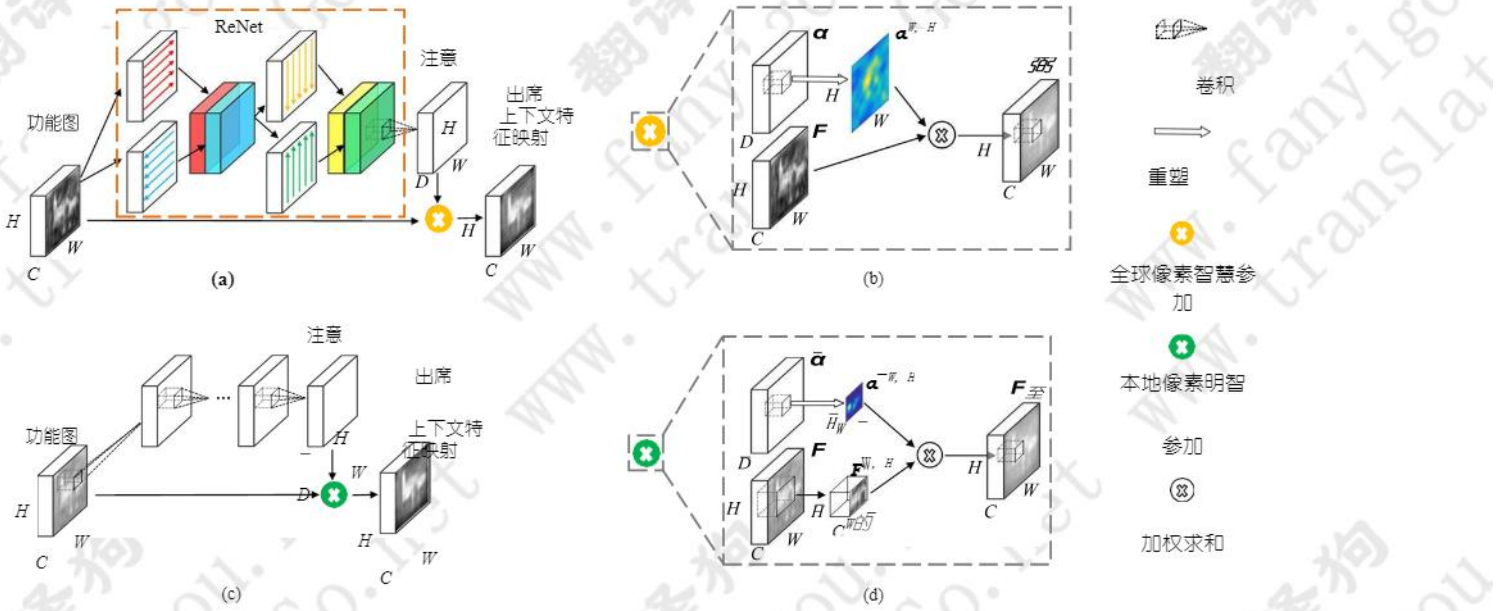


图2. (a) 提议的全球PiCANet的架构。 (b) 详细的全球参加活动的说明。 (c) 拟议本地PiCANet的架构。 (d) 详细说明当地的参加活动。

如图2 (d) 所示，对于像素 (w, h)，$F^{W, H}$中的特征通过加权相加$\alpha^{w, h}$得到F^{Ξ}:

$$F^{w, h} = \sum_{i=1}^{\tilde{D}} \alpha^{-W, H}_{i} f^{-W, H}_{i} \quad (3)$$

3.3. 有效和高效的实施

为了计算效率，所有像素的参与操作可以通过类似卷积的方式同时实现。 我们也可以采用洞算法[2]在参加的操作中，它支持稀疏的sam-通过使用扩张卷积来拍摄特征地图。 因此，我们可以使用扩展的小D或D 来关注大背景区域，以使PiCANets更高效。 PiCANets的梯度可以很容易地计算出来，通过反向传播算法可以实现端到端的训练[31]。 我们还可以在softmax标准化之前使用批量标准化 (BN) [11]层来使网络训练更有效。

4. 使用PiCANets进行显着对象检测

在本节中，我们详细介绍了采用PiCANets进行分层次显着检测的网络架构。 整个网络基于U-Net [30]架构，如图3 (a) 所示。 然而，与[30]不同的是，我们的U-Net的编码器是一个带有孔算法[2]的FCN来保持特征映射的分辨率。 解码器遵循U-Net的思想，使用跳过连接以及我们提出的嵌入式全局和本地PiCANets。

考虑到全球PiCANet需要输入特征图具有固定大小，我们设置输入图像

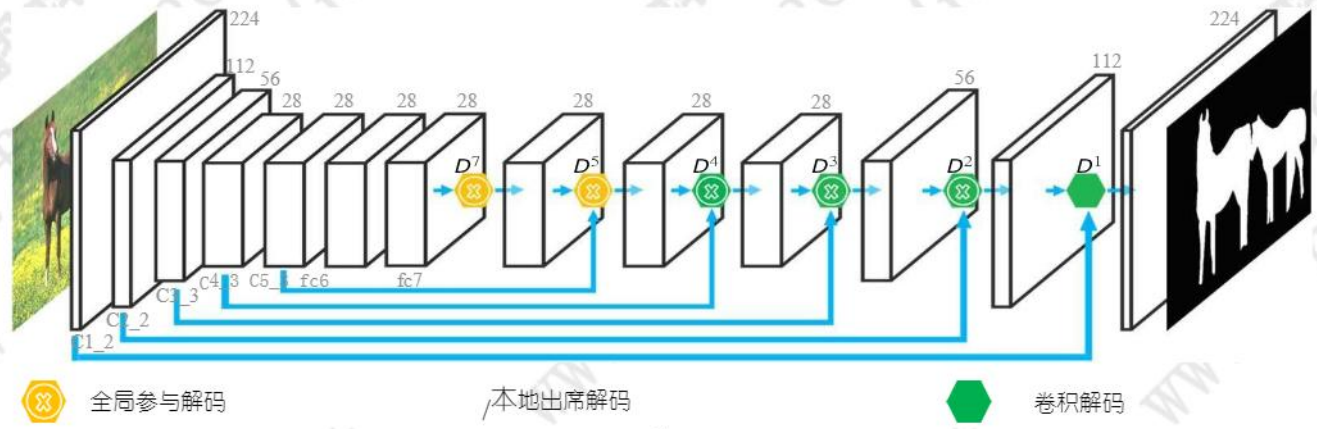
编码器部分是一个带有预训练骨干网络的FCN，例如VGG [33]网络或ResNet [8]。 我们以VGG 16层网络为例，其中包含13个Conv层，5个max-

合并图层和2个完全连接的图层。 如图3 (a) 所示，为了保留相对较大的空间大小

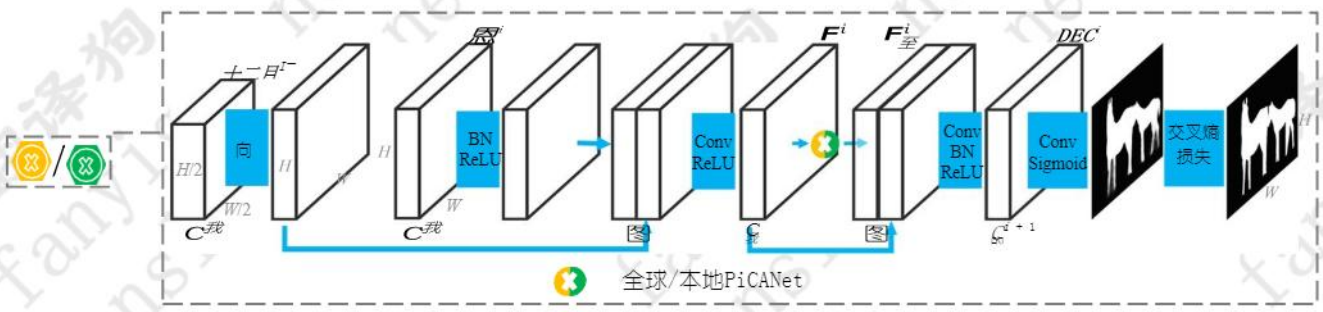
为了精确显着性检测，我们将pool4和pool5层的池化步长修改为1，并采用孔算法[2]为conv5层引入2的扩大。 我们也按照[2]将最后2个完全连接的层转换为Conv层。 具体而言，我们使用1024 3 3内核，fc6层扩展为12，fc7层扩展1024 1 1内核。 因此，整个编码器网络的步幅减小到8，并且空间尺寸减小

的最终特征地图是28×28。

接下来，我们阐述我们的解码器部分。 如图3 (a) 所示，解码器网络有6个解码模块，命名为$D^{7}, D^{5}, D^{4}, \ldots, D^{1}$。 如图3 (b) 所示，在^{i}中，其中i 7,5,4, ..., 1,我们通常通过融合大小为W的中间编码器特征映射En^{i}来生成解码特征映射Dec^{i} $H C^{i}$和前面的解码特征映射Dec^{i-1}，其大小为W / 2 $H/2$ C^{i}。 En^{i}是VGG编码器部分中的i Conv模块的ReLU激活之前的Conv特征映射，它们在图3 (a) 中标记。 我们首先使用BN层和En^{i}上的ReLU激活。 同时，我们上采样Dec^{i-1}以获得W的空间大小 H通过使用具有双线性插值的去卷积层。 接下来，我们连接这两个特征映射并通过使用Conv和ReLU层将它们融合到具有C^{i}通道的特征映射F^{i}中。 然后，我们利用F^{i}上的全局或本地PiCANet来获取其出席的上下文



(a)



(b)

图3. (a) 带有VGG 16层骨干的显着网络的体系结构。 我们只显示VGG网络的跳过连接的编码器层。 “C”表示“卷积”，而*表示解码模块。 空间尺寸标注在代表特征图的长方体上。 (b) 参加解码模块的插图。 En^{i}表示来自编码器的卷积特征映射网络。 Dec*表示解码特征映射。 F^{i}表示融合特征映射，F$^i_{至}$表示其出席的上下文特征映射。 “UP”表示上采样。 一些重要的空间大小和频道号码也被标记。

功能地图F$^i_{至}$。 最后我们将F^{i}和F$^i_{至}$ 进入Dec^i 与大小WHC^{i+1}，通过一个Conv层，一个BN层和一个ReLU层。 我们也通过深度监督来促进网络培训。具体来说，在每一个^{i}中，我们使用一个在Dec^{i}上具有S形活化的Conv层来生成一个大小为WH的显着图，然后使用调整大小的地面实况显着图来监督基于平均横截面的网络训练，熵损失。

在每个^{i}中，我们将C^{i}设置为与编码器网络中的第i个Conv块的信道编号相同。 我们在接下来的三个解码模块中采用全局PiCANets^{7}和^{5}以及本地PiCANets。 对于^1，我们简单地将En^{1}和Dec^{2}融合到Dec^{1}中，使用简单的Conv层来提高计算效率。 5.4节介绍了全局和本地PiCANets的不同嵌入选择的影响。

5. 实验

5.1. 数据集

我们使用六个广泛使用的显着性基准数据集来评估我们的方法。 SOD [29]包含300幅图像与复杂的背景和多个前景objects。 ECSSD [42]具有1,000个语义上有意义且复杂的图像。 PASCAL-S [21]数据集由从PASCAL VOC 2010分段数据集中选择的850个图像组成。 DUT-O [43]包括5,168个具有挑战性的图像，每个图像通常具有复杂的背景和一个或两个前景对象。 HKU-IS [18]包含4,447个低色彩对比度的图像和每幅图像中的多个前景物体。 最后一个是DUTS [36]数据集，它是当前最大的显着物体检测基准数据集。 它在训练集中包含10,553张图像，即DUTS-TR和测试集中的5,019张图像，即DUTS-TE。 大多数图像对显着性检测都具有挑战性的场景。

5.2. 评估指标

我们采用四个评估指标来评估我们的模型。 第一个是精确召回（PR）曲线。 具体来说，显着图首先被二值化，然后在变化的阈值下与地面实况进行比较，从而获得一系列精确召回值对来绘制PR曲线。

第二个度量是F-measure分数，com-

全面考虑精确度和召回率：

$$F_\beta = \frac{(1+\beta^2) Precision \times 召回}{\beta^2 Precision + 回顾}, \quad (4)$$

我们将β^2设置为0.3，如前面的工作所建议的那样。

但是，如[27]所示，传统的评估 -
这些指标很容易受到插值缺陷，依赖缺陷和同等重要性缺陷的影响，[illegible]衡量得分F_β来解决这些缺点。 我们也欢迎，

低[5,34,10]采用它作为我们的指标之一
默认设置在[27]中。

我们使用的第四个度量是平均绝对误差（MAE）。它计算预测显着图与相应的地面真实显着图之间的平均绝对每像素差异。

5.3. 实施细节

网络结构。 在解码模块中，图3（b）中的所有卷积核被设置为1 1.在每个全局PiCANet中，我们使用256个隐藏神经元作为ReNet，然后我们使用1 1 Conv层生成D = 100维注意力重量，可以重新塑造
10 10个关注地图。 在参加的活动中，我们使用扩展= 3来关注28 28全球环境。 在每个局部PiCANet中，我们首先使用7 7 Conv层，扩大= 2，零填充和ReLU激活来生成128个通道的中间特征映射。 然后
我们采用1×1的Conv层生成D = 49 dimen-

从中分辨出7个 7个关注地图
可以获得。 然后，我们利用这些地方关注地图来关注膨胀= 2和零填充的13 13个本地情景区域。

培训和测试。 我们遵循[38]和[36]中的建议，使用DUTS-TR集作为我们的训练集。 对于数据增强，我们只是简单地调整每个图像的大小
256 256随机镜像翻转和随机裁剪
224 224个图像区域进行训练。 整个网络使用带动量的随机梯度下降（SGD）进行端对端训练。 由于在每个解码模块中采用深度监督，我们凭经验将D^7, D^5, D^4, D^3, D^2中的损失分别按0.5,0.5,0.5,0.8,0.8和1进行加权调整。 我们从零开始对解码器部分进行训练，学习率为0.01，并以0.1倍的学习速率对编码器进行微调。 我们将批量设置为10，最大迭代步骤为20,000，并且每7000步将学习速率衰减0.1倍。 动量和重量衰减分别设定为0.9和0.0005。

我们基于Caffe [14]库实现我们的模型。

GTX Titan X GPU用于加速。 测试时，

每个图像被简单地调整大小为224×224，然后送入

表1.我们的模型和基线模型的不同设置的定量结果。 “MP”和“AP”分别表示最大池和平均池。 “+ 75G432LP”表示在D^7和D^5中使用全局PiCANet，在D^4，D^3，D^2中使用本地PiCANet。 其他设置可以被类似地推断。 蓝色表示最佳性能。

设置	DUT-O [43]			DUTS-TE [36]		
	F_β	F_β	MAE	F_β	F_β	MAE
U-Net [30]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
+75GP	0.778	0.662	0.071	0.834	0.724	0.057
+75G432LP	**0.794**	**0.691**	0.068	**0.851**	**0.748**	0.054
+MP	0.780	0.671	0.070	0.833	0.727	0.057
+AP	0.778	0.670	0.069	0.831	0.724	0.056
+75432LP	0.787	0.680	0.069	0.842	0.738	0.055
+7G5432LP	0.792	0.690	0.069	0.849	0.744	0.054
+754G32LP	0.794	0.688	**0.065**	0.850	0.747	**0.053**

该网络获取其显着图。 使用VGG 16层主干时，测试过程每个图像的成本仅为0.178s。 我们的代码将被发布。

5.4. 消融研究

拟议的PiCANets的有效性。 为了证明所提出的PiCANets的有效性，我们在表1中的两个具有挑战性的数据集上显示了我们的模型与基线模型的定量比较结果。 “U-Net”是没有PiCANets的基线网络。 “+ 75GP”的意思
我们只将两个全局PiCANet嵌入到D^7和D^5中，而“+ 75G432LP”表示我们将全局PiCANet嵌入到D^7
和D^5和本地PiCANets D^4，D^3，D^2。 com-
比较结果表明，当我们逐渐使用PiCANets选择性地结合全局和多尺度局部上下文时，模型性能可以逐步提高。 在补充材料中给出了更详细的逐步嵌入PiCANets在每个解码模块中的消融研究。

为了公平比较，我们还采用最大池（MP）和平均池（AP）来结合这些上下文。 表1显示，虽然使用这些非参数化池方案来整合全局和本地环境可以带来性能收益，但使用我们提出的PiCANets来选择信息环境是一种更好的方法。

我们还展示了视觉比较结果来展示提议的PiCANets的有效性。 在图5（a）中，我们展示了一幅图像及其地面真实显着图
(b)显示了基线U-Net（顶部）和我们的模型（底部）的预测显着图。 我们可以看到，我们的显着性模型可以在PiCANets的帮助下获得更一致的突出显着图。 在图5（c）中，我们显示了Conv特征映射F^5（顶部）与at-
倾向的上下文特征映射F^5 至 （下）与全球

PiCANet。 虽然（d）显示了F^2（顶部）至 （底部）
和F^2

当地的PiCANet。 我们可以看到全球的PiCANet

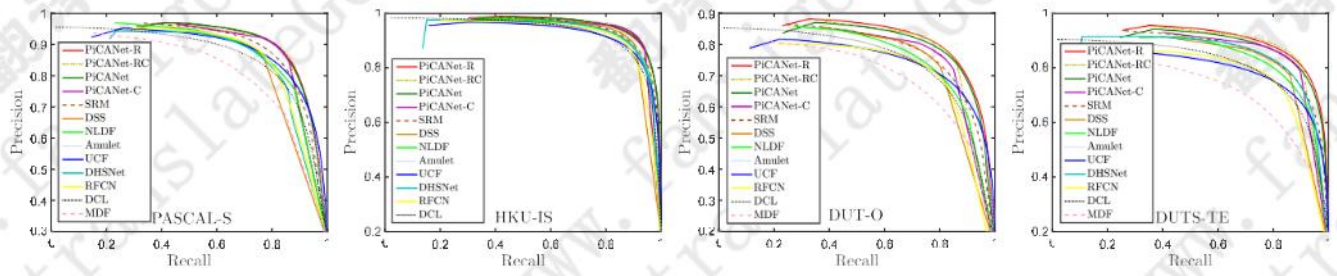


图4. 根据PR曲线对四个大数据集进行比较。

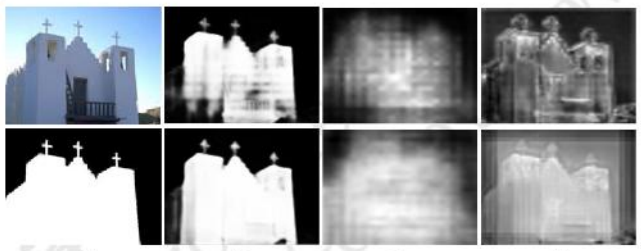


图5. 我们的模型与基线U-Net的视觉比较。 （a）图像及其基本事实。 （b）基线U-Net（顶部）和我们的模型（底部）的显着图。 （c）F^5（上）和

$F^5_{至}$（底部）。 （d）F^2（上至（底图）和F^2部）。

[5]有助于更好地区分前景对象和背景，而[2]中的局部PiCANet使特征映射更加均匀，这使得整个前景对象更加均匀地突出显示。

为了进一步理解为什么PiCANets可以实现这种改进，我们将图6中一个图像中两个像素的学习关注图可视化。在（b）列中，顶部图像显示背景像素的全局注意力主要关注前景对象而底部图像显示对于前景像素，它主要参与背景区域。 这一观察与全球对比机制非常匹配。 因此，我们的全球PiCANet可以帮助网络从背景中有效地分辨出显着的对象。 至于本地关注，由于我们对不同的解码模块使用了固定注意力大小（13 13），因此我们可以将多尺度关注从粗到细，大上下文合并为小尺寸，如图6中的红色矩形所示。（c ）和（d）列表明，局部注意力主要集中在具有相关像素的均匀区域，从而使显着图更加均匀，正如（a）栏底部图像所示。 更多的可视化可以在补充材料中找到。

嵌入选择的影响。 我们还显示了表1中我们的全球和本地PiCANets的不同嵌入选择的比较结果。它表明只嵌入本地PiCANets（“+ 75432LP”）较差。 而

图6. 提议的学习关注地图的插图
PiCANet。 （a）显示图像及其预测的显着图

我们的模型。 我们在[5]（b），‘（c）和[3]中展示了两个像素的关注图（用红色点表示，顶行表示背景像素，最下面的行表示前景像素）（d）。 出席的上下文区域由红色矩形标记。

“+ 7G5432LP”和“+ 754G32LP”的结果稍差于我们的最终选择，即“+ 75G432LP”。 我们不考虑在其他解码模块中使用全局PiCANets，因为ReNet对于大型特征映射是耗时的。

5.5. 与艺术级比较

我们将我们的显着性模型与其他9个最先进的模型比较，即SRM [38]，DSS [9]，NLDF [26]，护身符[45]，UCF [46]，DHS [23]，RFCN [37]，DCL [19]，和中密度纤维板[18]。

在表2中，我们显示了定量比较结果。 由于[19]和[9]采用完全连接的条件随机场（CRF）[16]作为后处理技术，而[38]使用ResNet50 [8]网络作为其主干，为了公平比较，我们也采用他们在我们的模型中，并与其他模型在不同的设置下进行比较。 图4给出了四个大数据集上的PR曲线。我们观察到，在所有设置下，我们的模型始终比所有其他模型执行得更好，尤其是在加权F-measure方面。 值得注意的是，即使只使用VGG 16层主干，也不需要任何后处理方法，我们的香草PiCANet仍然可以胜任所有其他型号。 使用两者时

表2. 不同设置下6个数据集上不同方法的比较。 蓝色表示每种设置下的最佳性能，而红色表示所有设置下的最佳性能。 “-C”，“-R”和“-RC”分别表示使用CRF后处理，ResNet50主干和两者。

数据集	SOD [43]			ECSSD [43]			PASCAL-S [21]			HKU-IS [18]			DUT-O [43]			DUTS-TE [36]		
公	F_β	F_β^ω	MAE	F_β	F_β^ω	MAE	F_β	F_β^ω	MAE	F_β	F_β^ω	MAE	F_β	F_β^ω	MAE	F_β	F_β^ω	MAE
VGG-16 [33]主干																		
密度板[18]	0.760	0.501	0.192	0.832	0.705	0.105	0.782	0.579	0.165	-	-	-	0.694	0.565	0.092	0.711	0.509	0.114
RFCN [37]	0.807	0.592	0.166	0.898	0.727	0.095	0.850	0.671	0.132	0.898	0.718	0.080	0.738	0.562	0.095	0.783	0.587	0.090
国土安全部[23]	0.827	0.686	0.133	0.907	0.841	0.060	0.841	0.732	0.111	0.902	0.806	0.054	-	-	-	0.829	0.698	0.065
UCF [46]	0.803	0.644	0.169	0.911	0.789	0.078	0.846	0.709	0.128	0.886	0.751	0.074	0.735	0.565	0.132	0.771	0.588	0.117
护身符[45]	0.808	0.686	0.145	0.915	0.841	0.059	0.858	0.762	0.103	0.896	0.813	0.052	0.743	0.626	0.098	0.778	0.657	0.085
NLDF [26]	0.842	0.708	0.130	0.905	0.839	0.063	0.845	0.743	0.112	0.902	0.838	0.048	0.753	0.634	0.080	0.812	0.710	0.066
PiCANet	**0.855**	**0.721**	**0.108**	**0.931**	**0.865**	**0.047**	**0.880**	**0.781**	**0.088**	**0.921**	**0.847**	**0.042**	**0.794**	**0.691**	**0.068**	**0.851**	**0.748**	**0.054**
VGG-16 [33]主干+ CRF [16]																		
DCL [19]	0.825	0.641	0.198	0.901	0.820	0.075	0.823	0.678	0.189	0.885	0.736	0.137	0.739	0.575	0.157	0.782	0.606	0.150
DSS [9]	**0.846**	0.718	0.126	0.916	0.871	0.053	0.846	0.751	0.112	0.911	0.866	0.040	0.771	0.691	0.066	0.825	0.754	0.057
PiCANet-C	0.836	**0.727**	**0.102**	**0.933**	**0.898**	**0.036**	**0.881**	**0.809**	**0.079**	**0.925**	**0.889**	**0.031**	**0.784**	**0.722**	**0.059**	**0.850**	**0.791**	**0.046**
ResNet50 [8]骨干网																		
SRM [38]	0.845	0.671	0.132	0.917	0.853	0.054	0.862	0.760	0.098	0.906	0.836	0.046	0.769	0.658	0.069	0.827	0.722	0.059
PiCANet-R	**0.858**	**0.723**	**0.109**	**0.935**	**0.867**	**0.047**	**0.881**	**0.780**	**0.087**	**0.919**	**0.840**	**0.043**	**0.803**	**0.695**	**0.065**	**0.860**	**0.756**	**0.051**
ResNet50 [8]主干+ CRF [16]																		
PiCANet-RC	0.856	**0.742**	**0.100**	**0.940**	**0.908**	**0.035**	**0.883**	**0.812**	**0.077**	**0.927**	**0.890**	**0.031**	**0.804**	**0.743**	**0.054**	**0.866**	**0.811**	**0.041**

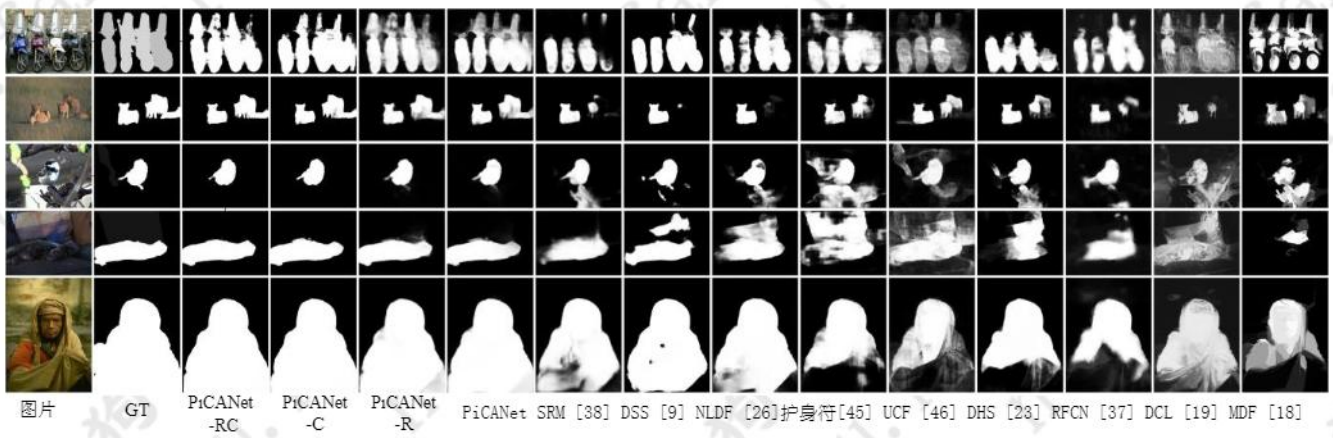


图7. 定性比较。 (GT：地面真相)

的CRF后处理和ResNet50骨干，我们的PiCANet-RC模型实现了最佳性能，并且与现有方法相比显示出显着的性能提升。

在图7中，我们展示了定性比较。 我们观察到，我们的模型可以处理各种具有挑战性的场景，包括具有复杂背景和前景（第1,2和3行）的图像，变化的物体比例，物体触摸图像边界（第5行），与背景具有相似外观的物体第4行）。 最重要的是，即使对于不使用任何后处理方法的香草PiCANet和PiCANet-R，它们也可以借助PiCANets比其他模型更加均匀地突出显示对象。 更多的视觉比较结果可以在补充材料中找到。

6. 结论

在本文中，我们提出了新颖的PiCANets来选择性地关注全局或本地上下文，并为每个像素构建信息性上下文特征。 我们应用PiCANets以分层方式检测显着对象。 借助参与的上下文，我们的模型在六个基准数据集上实现了最佳性能。 我们还提供对PiCANets的有效性的深入分析。

致谢

这项工作得到了国家科学基金（No. 61473231和61522207）和NSF CAREER（No. 1149783）的部分支持。

参考

[1] D. Bahdanau, K. Cho和Y. Bengio。 神经机器翻译通过联合学习来对齐和翻译。 在ICLR, 2015。2

[2] L.-C. 陈, G.帕潘德里欧, 科克金诺斯, K.墨菲和A. L.尤伊尔。 Deeplab：深度卷积网络的语义图像分割, 无限卷积和完全连接的crfs。 arXiv预印本arXiv：1606.00915,2016。4

[3] L.-C. Chen, Y. Yang, J. Wang, W. Xu和AL Yuille。 关注比例尺度：感知尺度的语义图像分割。 在CVPR, 2016。2

[4] M.-M. Cheng, NJ Mitra, X. Huang, PH Torr和S.-M.胡。 基于全局对比度的显着区域检测。 TPAMI, 37（3）：569-582,2015。1, 2

[5] C. Gong, D. Tao, W. Liu, SJ Maybank, M. Fang, K. Fu和J. Yang。 显着性从简单传播到困难。 在CVPR, 2015年。6

[6] A. Graves, N. Jaitly和A.-r. 穆罕默德。 具有深度双向lstm的混合语音识别。 在IEEE自动语音识别和理解研讨会上, 2013。3

[7] B. Han, H. Zhu和Y. Ding。 基于加权稀疏编码残差的自底向上显着性。 在ACM多媒体, 2011年。1

[8] K. He, X. Zhang, S. Ren和J. Sun. 图像识别的深度残留学习。 在CVPR, 2016。4,7,8

[9] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu和P. Torr。 深度监控短连接的显着物体检测。 在CVPR, 2017年1, 3, 7, 8

[10] P. Hu, B. Shuai, J. Liu和G. Wang。 用于显着物体检测的深层次集。 在CVPR, 2017.3,6

[11] S. Ioffe和C. Szegedy。 批量标准化：通过减少内部协变量来加速深度网络培训。 在ICML, 2015。4

[12] L. Itti, C. Koch和E. Niebur。 快速场景分析的基于显着性的视觉注意模型。 TPAMI, 20（11）：1254-1259,1998

[13] M. Jaderberg, K. Simonyan, A. Zisserman, 等人。 空间变压器网络。 在NIPS, 2015。3

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Gir-shick, S. Guadarrama和T. Darrell。 Caffe：用于快速特征嵌入的卷积体系结构。 在ACM Multimedia, 2014。6

[15] DA Klein和S. Frintrop。 显着物体检测的特征统计的中心 - 环绕发散。 在ICCV, 2011年。1, 2

[16] P.Kra henbu hl和V. Koltun。 高斯边缘电位的完全连接crfs的有效推断。 在NIPS, 2011.7,8

[17] J. Kuen, Z. Wang和G. Wang。 用于显着性检测的经常性注意网络。 在CVPR, 2016。1, 3

[18] G. Li和Y. Yu。 基于多尺度深度特征的视觉显着性。 在CVPR, 2015年。1, 3, 5, 7, 8

[19] G. Li和Y. Yu。 用于显着物体检测的深度对比学习。 在CVPR, 2016年。1, 3, 7, 8

[20] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan。 用于物体检测的细致背景。 IEEE Transactions on Multimedia, 19（5）：944-954, 2017。2

[21] Y.Li, X. Hou, C.Koch, JM Rehg和ALYuille。 显着物体分割的秘密。 在CVPR, 2014.5,8

[22] N. Liu和J. Han。 用于显着性检测的深度空间上下文长期循环卷积网络。 arXiv预印本arXiv：1610.01708, 2016

[23] N. Liu和J. Han。 Dhsnet：用于显着物体检测的深层次显着网络。 在CVPR, 2016年。1, 3, 7, 8

[24] N. Liu, J. Han, D. Zhang, S. Wen和T. Liu。 使用卷积神经网络预测眼部注意力。 在CVPR, 2015。1, 3

[25] J. Long, E. Shelhamer和T. Darrell。 用于语义分割的完全卷积网络。 在CVPR, 2015。1

[26] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li和P.-M. Jodoin。 用于显着物体检测的非局部深度特征。 在CVPR, 2017年1, 3, 7, 8

[27] R. Margolin, L. Zelnik-Manor和A. Tal。 如何评估前景地图？ 在CVPR, 2014年。6

[28] V. Mnih, N. Heess, A. Graves, 等人。 视觉注意力的复发模型。 在NIPS, 2014年。2

[29] V. Movahedi和JH Elder。 对显着物体分割的性能测量的设计和感知验证。 在CVPR研讨会上, 2010年5

[30] O. Ronneberger, P. Fischer和T. Brox。 U-net：用于生物医学图像分割的卷积网络。 在MIC-CAI, 2015.2,4,6

[31] DE Rumelhart, GE Hinton, RJ Williams, 等人。 通过反向传播错误学习表示。 认知建模, 5（3）：1, 1988

[32] P. Sermanet, A. Frome和E. Real。 注意细化分类。 arXiv预印本arXiv：1412.7054,2014。2

[33] K. Simonyan和A. Zisserman。 用于大规模图像识别的非常深的卷积网络。 arXiv预印本arXiv：1409.1556,2014。4,8

[34] 厕所。 Tu, S.He, Q.Yang和S.-Y. 简。 使用最小生成树的实时显着对象检测。 在CVPR, 2016年。6

[35] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville和Y. Bengio。 Renet：一种基于循环神经网络的卷积网络替代方案。 arXiv预印本arXiv：1505.00393,2015。3

[36] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X.阮。 学习使用图像级别监控来检测显着对象。 在CVPR, 2017.5,6,8

[37] L. Wang, L. Wang, H. Lu, P. Zhang和X. Ruan。 循环完全卷积网络的显着性检测。 在ECCV中, 2016年1, 3, 7, 8

[38] T. Wang, A. Borji, L. Zhang, P. Zhang和H. Lu。 用于检测图像中的显着对象的分阶段细化模型。 在ICCV, 2017年。1, 3, 6, 7, 8

[39] S. Xie和Z. Tu。 全局嵌套边缘检测。 在ICCV, 2015。3

[40] H. Xu和K. Saenko。 询问，参加并回答：探索以视觉问题回答为题的空间注意问题。 在ECCV，2016。2

[41] K. Xu，J. Ba，R. Kiros，K. Cho，A. Courville，R. Salakhudi-nov，R. Zemel和Y. Bengio。 显示，参加并讲述：神经图像标题生成与视觉注意力。 在ICML，2015。2

[42] Q. Yan，L. Xu，J. Shi和J. Jia。 分层显着性检测。在CVPR，2013。5

[43] C. Yang，L. Zhang，H. Lu，X. Ruan和M.-H. 杨。 通过基于图形的多方排名进行显着性检测。 在CVPR，2013年。2,5,6,8

[44] Z. Yang，X. He，J. Gao，L. Deng和A. Smola。 堆叠的图像问题回答网络。 在CVPR，2016。2

[45] P. Zhang，D. Wang，H. Lu，H. Wang和X. Ruan。 护身符：为显着物体检测汇总多级卷积特征。 在ICCV，2017年1，3，7，8

[46] P. Zhang，D. Wang，H. Lu，H. Wang，and B. Yin。 学习不确定的卷积特征用于精确显着性检测。 在ICCV，2017年1，7，8

[47] R. Zhao，W. Ouyang，H. Li，and X. Wang。 多情境深度学习的显着性检测。 在CVPR，2015。1，3

PiCANet：针对显着性检测补充材料学习像素式上下文关注

年刘[1] 韩俊伟[1] 杨明轩[2,3]
[1]西北工业大学 [2]加州大学默塞德分校 [3]Google Cloud
{liunian228, junweihan2010}@gmail.com mhyang@ucmerced.edu

在这个补充材料中，我们包含更多的实施细节，更多的消融分析和更多的实验结果。

1. 其他实施细节

1.1. 使用ResNet50骨干网

当使用ResNet50网络[1]作为主干时，我们修改4个^{日}和5个^{日}残留块，使其步幅分别为1和2和4的扩大，从而使编码器的步幅为8然后我们逐步融合解码模块^{5}中的5^{日}到1^{57}Conv模块的特征映射到^{1}。 我们采用全球PiCANets
D ^{5}和^{4}，以及最后三个D模块中的本地PiCANets。 在每个解码模块^{i}中，我们使用ResNet50编码器中的i^{日}Conv模块的最终Conv特征图（例如res4f和res3d）作为并入的编码器特征映射En^{i}，并且不采用BN其上的ReLU层如图3（b）所示，因为ResNet50网络已经在每个Conv层之后使用了BN层。 由于conv1图层的步幅为2，因此最终生成的显着图大小为112 112。

与使用VGG-16主干时一样，我们根据经验将^{5}，^{4}，^{1}中的损耗权重设置为0.5，
D D . . . D
分别为0.5，0.8，0.8和1。 我们的小批量尺寸由于GPU内存限制，基于ResNet50的网络设置为8。 其他超参数设置与基于VGG-16的网络中使用的参数相同。一张图像的测试时间为0.236秒。

1.2. 使用CRF后处理

当我们采用CRF后处理方法时，我们使用[2]中使用的相同参数和相同代码。 另外，每张图片还需要0.09秒的费用。

表1. 逐步嵌入PiCANets的有效性。 “+ 75G432LP”表示在^{7}和^{5}中使用全局PiCANet，在^{3}，^{2}中使用本地PiCANet。 其他设置可以被类似地推断。 蓝色表示最佳性能。

设置	DUT-O [11]			DUTS-TE [8]		
	F_β	F_β^w	MAE	F_β	F_β^w	MAE
U-Net [7]	0.761	0.651	0.073	0.819	0.715	0.060
+7GP	0.772	0.660	0.071	0.826	0.722	0.058
+75GP	0.778	0.662	0.071	0.834	0.724	0.057
+75G4LP	0.785	0.678	0.069	0.840	0.736	0.056
+75G43LP	0.791	0.682	**0.068**	0.848	0.740	0.055
+75G432LP	**0.794**	**0.691**	**0.068**	**0.851**	**0.748**	**0.054**

2. 实验

2.1. 逐步嵌入Pi-Canets的有效性

在这里，我们报告了一个更详细的消融研究，逐步在每个解码模块中嵌入PiCANets。 如表1所示，在TF699，TF700，TF701，TF702中逐步嵌入全局和局部PiCANets可以一致地提高显着性检测性能，从而证明我们提出的PiCANets和显着性检测模型的有效性。
D D D D

2.2. 学习注意地图的更多可视化

我们在图1中为五个参与解码模块说明了更多的学习关注地图。 图1显示在^{7}和^{5}中学习的全局关注可以用于前景对象的背景像素和前景像素的背景D。 在^{4}，^{3}和^{2}中学习到的本地关注可以关注与所涉像素具有相似语义的区域。

2.3. 我们的模型和Stata最先进的方法之间的更多视觉比较

D D

我们在图2中也显示出更多的定性结果。它表明，与其他最先进的方法相比，我们的模型可以更准确地突出显示突出物体，

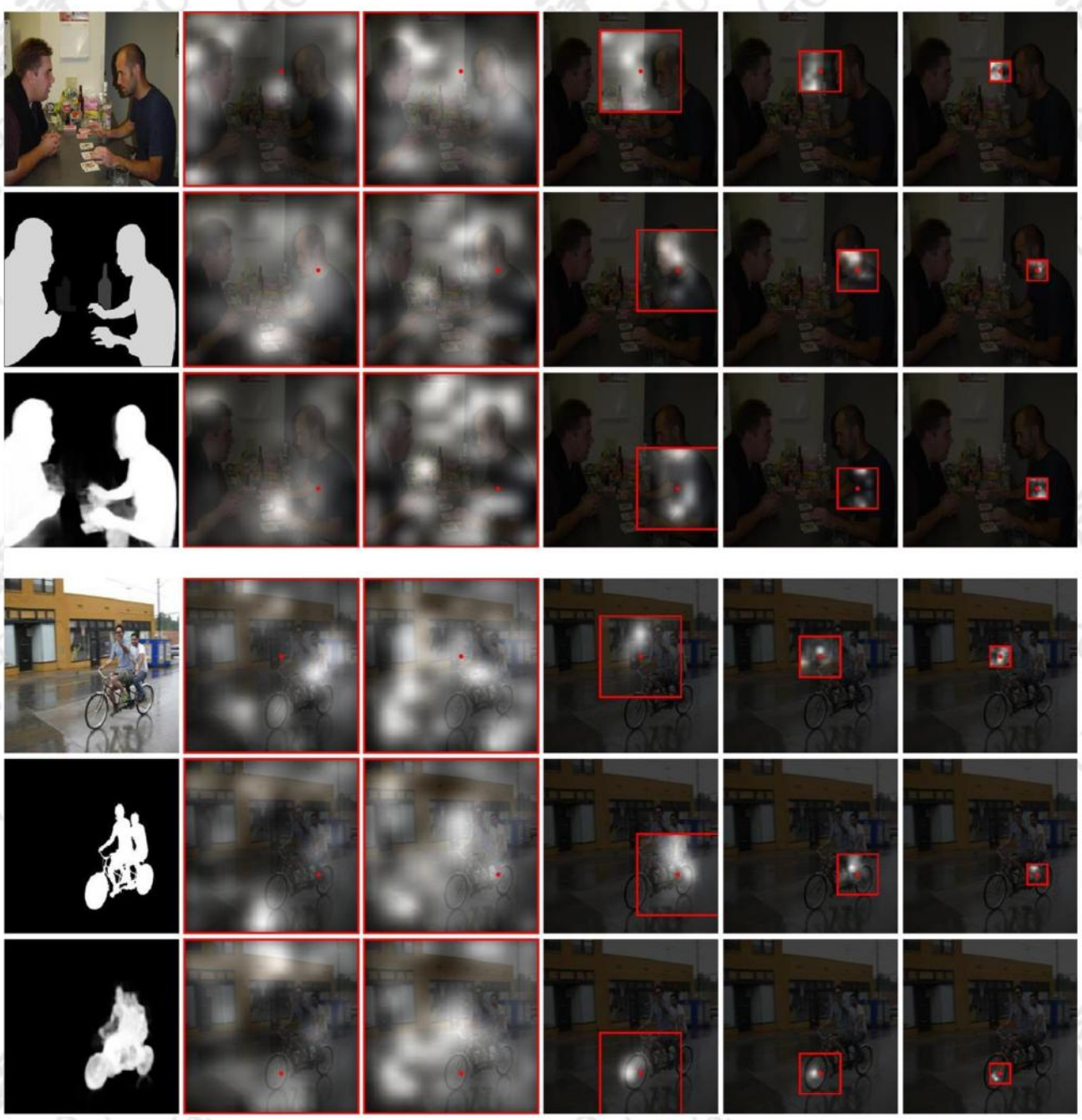


图1. 提议的PiCANets的学习关注地图的插图。 第一列显示了两个图像及其相应的地面实况蒙版和我们模型的预测显着图，而最后五列分别显示了五个参与解码模块中的关注图。 对于每幅图像，我们给出三个示例像素（用红色点表示），第一行显示背景像素，最下面两行显示两个前景像素。 出席的上下文区域由红色矩形标记。

即使不使用后处理技术，也可以在各种具有挑战性的情况下统一进行。

2.4. 失败案例

我们在图3中显示了我们的PiCANet-R模型的一些失败案例。基本上，当图像没有明显的前景对象时，我们的模型通常会失败，如（a）所示，

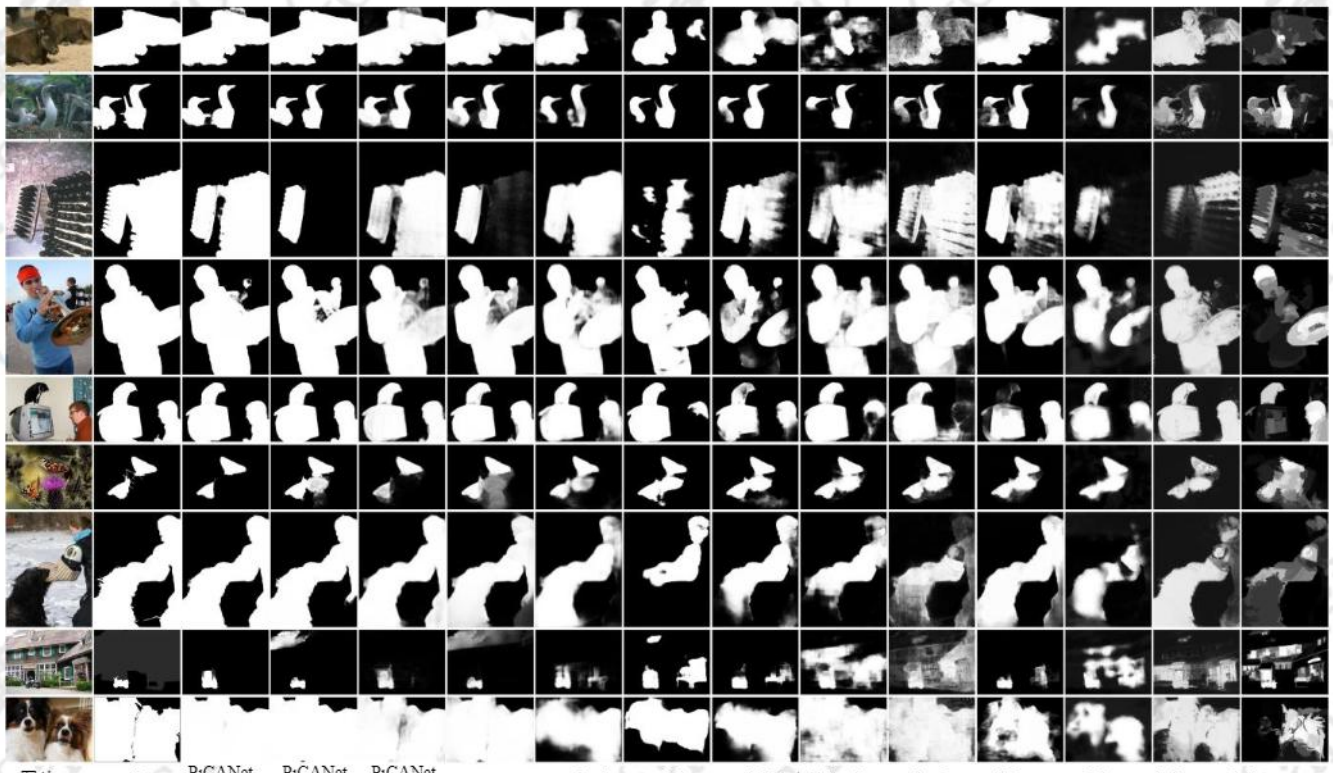


图2. 定性比较。 （GT：地面真相）

图3. 失败案例。 每组中的三幅图像分别是输入图像，地面实况和我们的结果。

和（b）。 （c）显示，当前景物体非常大时，我们的模型也很容易失败。 虽然这两种情况对其他传统和基于深度学习的显着模型也具有挑战性，但表明我们仍有很大空间来改进现有模型。 （d）显示物体上的不均匀照明也可能误导我们的模型。

参考

[1] K. He，X. Zhang，S. Ren和J. Sun. 图像识别的深度残留学习。 在CVPR，2016年。1

[2] Q. Hou，M.-M. Cheng，X. Hu，A. Borji，Z. Tu和P. Torr。 深度监控短连接的显着物体检测。 在CVPR，2017。1，3

[3] G. Li和Y. Yu。 基于多尺度深度特征的视觉显着性。 在CVPR，2015。3

[4] G. Li和Y. Yu。 用于显着物体检测的深度对比学习。 在CVPR，2016。3

[5] N. Liu和J. Han。 Dhsnet：用于显着物体检测的深层次显着网络。 在CVPR，2016。3

[6] Z. Luo，A. Mishra，A. Achkar，J. Eichel，S. Li和P.-M. Jodoin。 用于显着物体检测的非局部深度特征。
在CVPR，2017。3

[7] O. Ronneberger，P. Fischer和T. Brox。 U-net：用于生物医学图像分割的卷积网络。 在MIC-

CAI, 2015。1

[8] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X.阮。 学习使用图像级别监控来检测显着对象。 在CVPR, 2017年。1

[9] L. Wang, L. Wang, H. Lu, P. Zhang和X. Ruan。 循环完全卷积网络的显着性检测。 在ECCV, 2016。3

[10] T. Wang, A. Borji, L. Zhang, P. Zhang和H. Lu。 用于检测图像中的显着对象的分阶段细化模型。 在ICCV, 2017。3

[11] C. Yang, L. Zhang, H. Lu, X. Ruan和M.-H. 杨。 通过基于图形的多方排名进行显着性检测。 在CVPR, 2013年。1

[12] P. Zhang, D. Wang, H. Lu, H. Wang和X. Ruan。 护身符:为显着物体检测汇总多级卷积特征。 在ICCV, 2017。3

[13] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin。 学习不确定的卷积特征用于精确显着性检测。 在ICCV, 2017。3