基于多模态特征融合监督的RGB-D图像显著性检测

刘政怡* 段群涛 石 松 赵 鹏
(安徽大学计算机科学与技术学院 合肥 230601)

摘 要：RGB-D图像显著性检测是在一组成对的RGB和Depth图中识别出视觉上最显著突出的目标区域。已有的双流网络，同等对待多模态的RGB和Depth图像数据，在提取特征方面几乎一致。然而，低层的Depth特征存在较大噪声，不能很好地表征图像特征。因此，该文提出一种多模态特征融合监督的RGB-D图像显著性检测网络，通过两个独立流分别学习RGB和Depth数据，使用双流侧边监督模块分别获取网络各层基于RGB和Depth特征的显著图，然后采用多模态特征融合模块来融合后3层RGB和Depth高维信息生成高层显著预测结果。网络从第1层至第5层逐步生成RGB和Depth各模态特征，然后从第5层到第3层，利用高层指导低层的方式产生多模态融合特征，接着从第2层到第1层，利用第3层产生的融合特征去逐步地优化前两层的RGB特征，最终输出既包含RGB低层信息又融合RGB-D高层多模态信息的显著图。在3个公开数据集上的实验表明，该文所提网络因为使用了双流侧边监督模块和多模态特征融合模块，其性能优于目前主流的RGB-D显著性检测模型，具有较强的鲁棒性。

关键词：RGB-D显著性检测；卷积神经网络；多模态；监督
中图分类号：TP391.41 **文献标识码**：A **文章编号**：1009-5896(2020)04-0997-08
DOI: 10.11999/JEIT190297

RGB-D Image Saliency Detection Based on Multi-modal Feature-fused Supervision

LIU Zhengyi DUAN Quntao SHI Song ZHAO Peng
(*School of Computer Science and Technology, Anhui University, Hefei* 230601, *China*)

Abstract: RGB-D saliency detection identifies the most visually attentive target areas in a pair of RGB and Depth images. Existing two-stream networks, which treat RGB and Depth data equally, are almost identical in feature extraction. As the lower layers Depth features with a lot of noise, it causes image features not be well characterized. Therefore, a multi-modal feature-fused supervision of RGB-D saliency detection network is proposed, RGB and Depth data are studied independently through two-stream , double-side supervision module is used respectively to obtain saliency maps of each layer, and then the multi-modal feature-fused module is used to later three layers of the fused RGB and Depth of higher dimensional information to generate saliency predicted results. Finally, the information of lower layers is fused to generate the ultimate saliency maps. Experiments on three open data sets show that the proposed network has better performance and stronger robustness than the current RGB-D saliency detection models.

Key words: RGB-D saliency detection; Convolutional Neural Network(CNN); Multi-modal; Supervision

1 引言

图像显著性检测的目的是通过智能算法模拟人的视觉特点，提取图像中人类更加关注的区域。其在各种计算机视觉任务中具有广泛应用前景，例如图像检索[1]、图像压缩[2]、视觉追踪[3]等。近年来越来越多的显著检测研究工作设计了大量的深度卷积神经网络(Convolutional Neural Networks, CNNs)用于RGB显著性目标检测，并取得了较好的性能。然而，当显著目标和背景相似时，这些RGB显著性检测模型可能无法区分显著目标与背景。事实上，与提供了详细外观和纹理信息RGB数据相比，Depth数据包含了清晰的目标形状及丰富的空间结构，可以提供很多额外的显著性线索。此外深度传感器(例如微软Kinect或英特尔Real-

收稿日期：2019-04-29；改回日期：2019-08-31；网络出版：2019-09-05
*通信作者： 刘政怡 liuzywen@ahu.edu.cn
基金项目：安徽省自然科学基金(1908085MF182)，国家自然科学基金(61602004)，安徽高校自然科学研究项目(KJ2019A0034)
Foundation Items: The Provincial Natural Science Foundation of Anhui (1908085MF182), The National Natural Science Foundation of China (61602004), The Anhui University Natural Science Research Project (KJ2019A0034)

Sense)对光照变化的感知鲁棒性，大大有助于扩展显著性检测的应用场景。因此，对于RGB-D显著性检测任务，如何充分融合RGB和Depth信息是关键问题。

针对如何清晰形成RGB和Depth两种模态之间的互补信息，并将其充分融合，大多数先前的探索RGB和Depth数据跨模态互补的RGB-D融合网络分为单流网络架构[4]和双流网络架构[5–8]两种。单流网络架构认为Depth数据可作为无差别通道与RGB数据串联，通过网络共同学习RGB和Depth特征来得到显著图。文献[4]对输入的RGB数据和Depth数据进行超像素分割后，计算每一个超像素区域的显著特征向量，然后将计算好的显著值特征向量作为网络的输入，由网络结合超像素的显著性特征将RGB信息和Depth信息紧密耦合生成显著图。双流网络架构通过两个独立流分别学习RGB数据和Depth数据，然后通过一个在早期或晚期添加的共享网络层学习RGB和Depth特征的联合表示，来获取最终显著图。文献[5]将RGB数据和Depth数据作为两个小网络输入，单独训练，然后将生成的RGB和Depth特征通过多路径、多模态交互组成一个融合网络，共同训练。文献[6]提出一种基于CNN的框架来自动融合RGB和Depth数据从而获取显著图。文献[7]中提出一种晚融合网络模型捕捉RGB和Depth两种模态的高阶特征来生成显著图。文献[8]认为这种只融合RGB和Depth模态的深度CNN特征，交叉模态的互补信息不太可能被很好地捕获。因此，提出一个渐进互补感知的融合网络来有效利用多层次的跨模态互补信息。人们普遍认为，不同层次的特征是互补的，它们在不同的尺度上对场景进行抽象。然而，不是所有层次的跨模态信息都是互补的。

因此，针对如何融合不同层次的跨模态信息问题，本文采用双流网络结构。首先将RGB图和Depth图分别作为两个VGG16Net[9]的网络输入，采用双流侧边监督模块对RGB流和Depth流进行显著预测监督，加快网络收敛速度，帮助网络更好地学习各层特征。为了充分地利用和融合网络的不同层次的RGB和Depth的语义信息，对网络采取高层指导低层、从全局到局部的方式得到最终显著预测结果。其中，对网络后3层的高维多模态信息构造了多模态特征融合模块生成多尺度多模态融合特征，以此得到网络高层显著预测结果，同时由于网络低层特征包含目标细节信息，对网络前两层特征融合时不采用多模态特征融合模块，特别地，在实验中发现位于低维的Depth特征并不好，导致最终显著预测结果会出现噪声。为了消除低维Depth特征带来的负面效果，本文选择在低层特征融合中不加入低维Depth特征。在广泛使用的数据集上的实验表明本文模型优于目前主流的RGB-D显著性检测模型，具有较强的鲁棒性。

2 本文方法

本文方法使用了两个VGG16Net作为主干基础网络，如图1所示。用RGB图和Depth图分别作为输入，提取$m\{m=1,2,\cdots,5\}$层的RGB和Depth特征，形成RGB流和Depth流。由于网络高层特征获取的是显著目标的高维语义信息，而忽略了目标的边界信息。因此本文采取了高层指导低层，由深到浅、从全局到局部，分别获取各层基于RGB和Depth特征的显著图及多模态融合显著图，在真值图的监督指导下优化网络参数，最终以RGB流的显著图输出$P^{1r}\left(P^{f}\right)$作为最终预测结果。其中，对网络第$m\{m=3,4,5\}$层，将RGB和Depth特征串

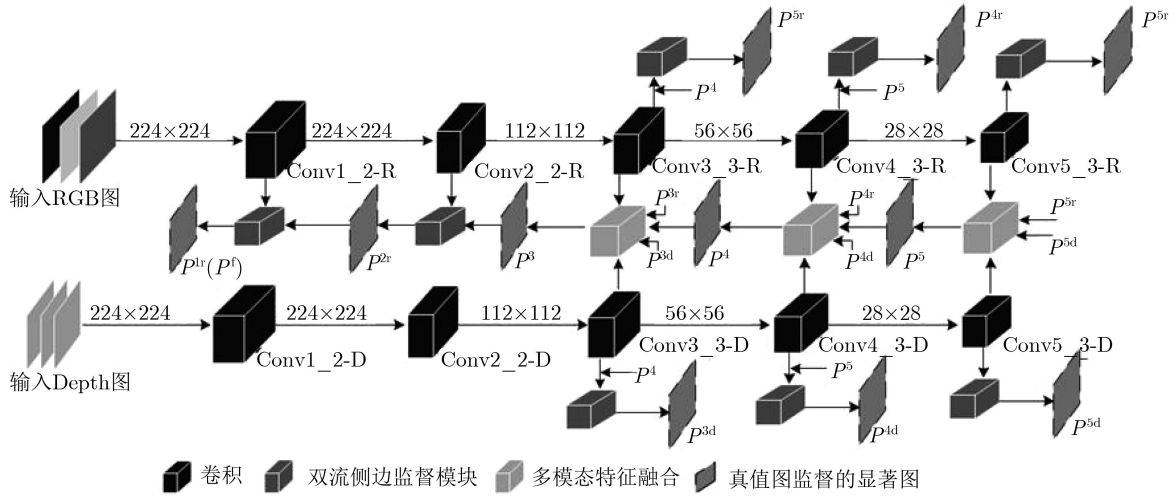


图 1 本文方法模型

联，其侧边经过真值图监督的显著输出和上一层的显著输出结果作为指导，使用多模态特征融合模块得到高层显著预测结果。对网络第$m\{m=1,2\}$层，由于低层特征更多关注的是局部信息，故在前两层特征融合中不使用高层特征融合方式。且低层Depth信息不好影响最终预测结果，故将低两层的depth流监督删除，每一层的显著图都受上一层的指导。

2.1 双流侧边监督

根据文献[10,11]可以得出网络监督可以促进网络收敛速度并且生成更好的层次表征以满足各阶段特征需求的结论。考虑到单卷积会使通道数骤降而损失较多信息，故使用3次卷积逐步降低特征通道数的方法，同时文献[11,12]充分考虑到深层高维特征输出保留更多的是目标信息和位置信息而忽略了目标细节信息，而低维特征更多关注局部信息和边界信息，也就是说高层显著图对于显著物体检测有着更好的指导作用，因此，对VGG16Net主干基础网络每个层特征经过3次卷积操作后通道数降为64，然后结合高一层输出的显著图P^{m+1}，再通过卷积、反卷积及卷积操作，产生本层显著图。在网络学习过程中用真值图去监督，这样可以帮助网络更好地学习到各层特征。以上操作针对RGB流和Depth流分别处理，如图2所示，具体用公式表示为

$$\begin{aligned}P^{m\mathrm{r}}&=\delta\big(\mathrm{Conv}\left(\mathrm{Dec}\left(\mathrm{Conv}\left(3\mathrm{Conv}\left(R^m\right)\odot P^{m+1}\right)\right)\right)\big)\\P^{m\mathrm{d}}&=\delta\big(\mathrm{Conv}\left(\mathrm{Dec}\left(\mathrm{Conv}\left(3\mathrm{Conv}\left(D^m\right)\odot P^{m+1}\right)\right)\right)\big)\end{aligned}\tag{1}$$

其中，R^m和D^m分别表示第$m\{m=1,2,\cdots,5\}$层的RGB特征和Depth特征，P^{m+1}表示高一层输出的显著图，$\odot$表示按通道维度的串联操作表积操作，Dec(•)表示去卷积操作，目的是将特征大小变为与输入图片大小一致且其维度为1，而δ为Sigmoid函数来生成显著输出$P^{m\mathrm{r}}$, $P^{m\mathrm{d}}$。

依据文献[13,14]及本文实验可以得出前两层的Depth特征的可信度不高的结论，也就是说，包含更多局部信息的Depth特征对于整张RGB-D图像显著目标检测没有起到正面的作用，因此，取消了对Depth流中前两层特征的提取与监督。

同时，在实验过程中也发现P^{m+1}所表示的高层显著图取高层RGB流和Depth流融合后的显著图时，效果更佳，因此，在图1右侧第1行和第5行中，双流的高3层的侧边特征计算都使用的是融合后的显著图，特别地，在计算第5层的$P^{m\mathrm{r}}$和$P^{m\mathrm{d}}$没有使用高层P^{m+1}的指导。

2.2 多模态特征融合

考虑到网络高层产生的特征具有完整的关键信息，而单纯的使用一个尺度的卷积操作生成显著图，可能会将某些坏的特征图中的噪声无限制传递到显著预测输出中。因此本文对主干网络VGG16Net的第$m\{m=3,4,5\}$层的RGB和Depth特征，提出一种多模态特征融合方法，如图3，将两种特征串联，经过卷积操作将特征通道成倍缩小。

然而缺乏高层信息或者输出的指导，直接通过监督方式优化的侧边输出与真值图之间的误差会变大而且其效果也不佳。同双流侧边监督模块相似，用高层输出P^{m+1}和经过监督的RGB流输出$P^{m\mathrm{r}}$和Depth流输出$P^{m\mathrm{d}}$为多模态特征提供语义信息和位置信息，加快网络收敛速度，优化目标边界，得到更合适的多模态融合特征F^m，具体计算方法为

$$\begin{aligned}F^m=&\delta(\mathrm{Conv}(R^m\odot D^m))\odot\delta(\mathrm{Conv}(P^{m+1}))\\&\odot\delta(\mathrm{Conv}(P^{m\mathrm{r}}))\odot\delta(\mathrm{Conv}(P^{m\mathrm{d}}))\end{aligned}\tag{2}$$

其中，δ表示激励函数Sigmoid，将特征值和显著值归一化到同一个区间防止显著输出图被忽略[15]。RGB和Depth多模态特征融合，相比于分别处理单RGB特征和单Depth特征的方式，可以将鲁棒的层次特征表征进行跨模态信息互补与融合，为生成基于多模态特征的层次输出作铺垫。在形成多模态融合特征F^m之后，本文利用多尺度卷积模块[16]挖掘更强的融合特征表示。多尺度卷积模块提取多尺度上下文信息，其目的是获取一个空间响应映射，从而自适应地对每个位置的特征映射进行加权，通过为每个像素学习权重，使每个给定的输入定位最关注的部分，从而更适用于背景复杂的场景。多尺度卷积模块采用4个尺度的卷积核(1×1, 3×3, 5×5, 7×7)，不同卷积核大小的卷积层有着不同大小的感受野，可以获取不同尺度的特征信息。同时由于更大的卷积核对应更多的网络参数，本文对文献[16]提出的多尺度卷积模块进行修改，通过在5×5的卷

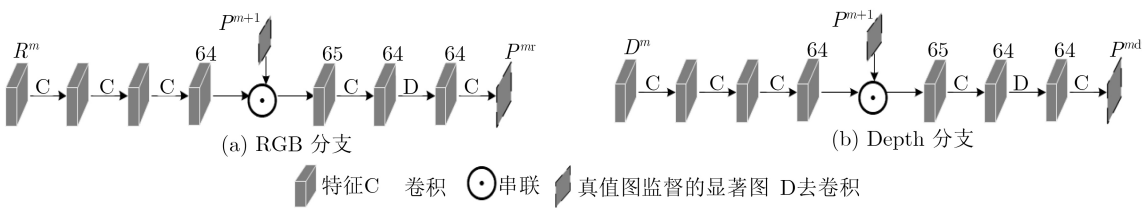


图 2 双流侧边监督模块

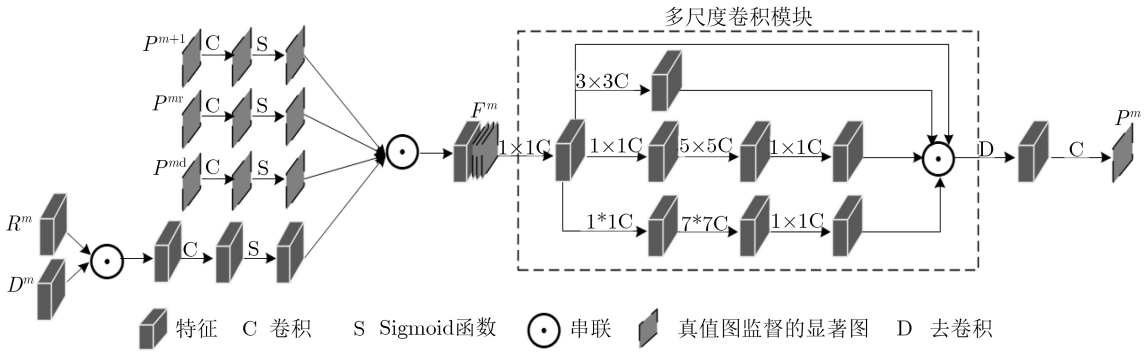


图 3 多模态特征融合方法

积层和7×7的卷积层前后增加1×1的卷积层来降低通道数而后恢复通道数的方式降低网络参数，然后通过通道串联方式形成多尺度多模态融合特征f_{cat}^m。

$$\begin{aligned} f_{\text{Ini}}^m &= \text{Conv}_{1\times1}(F^m) \\ f_{\text{cat}}^m &= f_{\text{Ini}}^m \odot \text{Conv}_{3\times3}(f_{\text{Ini}}^m) \odot \text{Conv}_{1\times1}(\text{Conv}_{5\times5}(\text{Conv}_{1\times1}(f_{\text{Ini}}^m))) \\ &\quad \odot \text{Conv}_{1\times1}(\text{Conv}_{7\times7}(\text{Conv}_{1\times1}(f_{\text{Ini}}^m))) \end{aligned} \tag{3}$$

则对应的多尺度多模态融合特征的显著输出P^m为

$$P^m = \delta(\text{Conv}(\text{Dec}(f_{\text{cat}}^m)) \tag{4}$$

主干网络VGG16Net第$m\{m=3,4,5\}$层的多模态融合特征更好地结合了RGB和Depth的高维特征，并通过多尺度卷积的处理后，对显著物体特征表征得更好，而第$m\{m=1,2\}$层学习的特征更多的是颜色、边缘等低维信息，并不适合使用多尺度卷积操作。因为多尺度卷积对应不同的卷积核，卷积核越大，其对应的感受野越大，看到的全局信息越多，低维特征保留更多的是目标细节信息，大的卷积核可能破坏其完整性。因此本文在前两层低维语义信息融合部分不采用高层融合中多尺度卷积模块处理融合特征的方式。融合网络产生的显著图同样受真值图监督，以学习到更好的多模态融合特征。正如2.1节所述，低层Depth信息可信度不高，会影响最终结果，因此本文将低两层的Depth流监督删除，每一层的显著图都受上一层的指导，最终取RGB流的显著图输出$P^{1r}(P^f)$为最终预测结果。

2.3 损失函数

对于VGG16Net的5个层，都有其对应的损失函数，可以加快收敛速度，获取特定层次表征的特征信息，并且可以缓解在训练过程中梯度消失问题，如图1所示。

给定训练集$T=\{(X_{\text{r}}^n, X_{\text{d}}^n, Y^n)|n=1,2,\cdots,N\}$，其中$X^n=\{x_i^n|i=1,2,\cdots,C\}$表示第$n$张输入RGB图和Depth图，$Y^n=\{y_i^n|i=1,2,\cdots,C, y_i^n\in[0,1]\}$表示相应的显著图，$N$表示训练集的张数，$C$表示一张图片中像素的个数。本文使用一个标准的交叉熵损失函数计算输入图片和真值图片之间的误差。对应于RGB流、Depth流和多模态特征的损失函数分别为

$$\begin{aligned} &L_{\text{side}}^{\text{mr}}(W^{\text{r}}, w^{m\text{r}}) \\ &= -\sum_i^C (y_i^n \lg P_{\text{r}}(y_i^n=1|x^{\text{r}}, W^{\text{r}}, w^{m\text{r}}) \\ &\quad + (1-y_i^n)\lg P_{\text{r}}(y_i^n=0|x^{\text{r}}, W^{\text{r}}, w^{m\text{r}})) \end{aligned} \tag{5}$$

$$\begin{aligned} &L_{\text{side}}^{\text{md}}(W^{\text{d}}, w^{m\text{d}}) \\ &= -\sum_i^C (y_i^n \lg P_{\text{r}}(y_i^n=1|x^{\text{d}}, W^{\text{d}}, w^{m\text{d}}) \\ &\quad + (1-y_i^n)\lg P_{\text{r}}(y_i^n=0|x^{\text{d}}, W^{\text{d}}, w^{m\text{d}})) \end{aligned} \tag{6}$$

$$\begin{aligned} &L_{\text{side}}^{m}(W^{\text{r}}, W^{\text{d}}, w^{\text{m}}) \\ &= -\sum_i^C (y_i^n \lg P_{\text{r}}(y_i^n=1|x^{\text{r}}, x^{\text{d}}, W^{\text{r}}, W^{\text{d}}, w^{m}) \\ &\quad + (1-y_i^n)\lg P_{\text{r}}(y_i^n=0|x^{\text{r}}, x^{\text{d}}, W^{\text{r}}, W^{\text{d}}, w^{m})) \end{aligned} \tag{7}$$

则最终的损失函数L_{cost}为

$$\begin{aligned} L_{\text{cost}} = &\sum_{m=1}^M L_{\text{side}}^{\text{mr}}(W^{\text{r}}, w^{m\text{r}}) + \sum_{m=3}^M (L_{\text{side}}^{\text{md}}(W^{\text{d}}, w^{m\text{d}}) \\ &+ L_{\text{side}}^{\text{md}}(W^{\text{r}}, W^{\text{d}}, w^{m})) \end{aligned} \tag{8}$$

其中，$W^{\text{r}}, W^{\text{d}}, w^{m\text{r}}, w^{m\text{d}}, w^m$分别表示RGB主干网络权值，Depth主干网络权值，RGB流侧边权值，Depth流侧边权值和多模态侧边网络权值，$M=5$。

本文在训练阶段通过优化式(9)损失函数更新权值

$$(W^{\text{r}}, W^{\text{d}}, w^{m\text{r}}, w^{m\text{d}}, w^m)^* = \arg\ \min L_{\text{cost}} \tag{9}$$

3 实验结果

3.1 数据集

本文在3个最广泛使用的数据集上评估本文模型。NLPR1000[17]数据集包含1000张RGB图和Depth图和与之对应的真值图，包含11种室内和室

外场景，超过400种物体。NJU2000[18]包含2003张立体RGB图像和对应的手工标注的真值图，其Depth图由光流法生成。STEREO[19]数据包含797张RGB图像和对应的真值图(GT)，这些RGB图像主要从英特网和3D电影中收集得到，其Depth图由光流法生成。为了公平对比，和文献[8]相似，采用其相同的训练集和测试集进行训练和评估。为了解决训练集不足的问题，本文对训练集进行了数据增强操作，即对原始图片进行翻转和边界1/10裁剪操作，以保留主要目标信息，训练集增加了16倍。

3.2 评估标准

评估标准被用于评价不同显著目标检测方法的性能。本文中采用5种评估准则评价模型和其他模型的好坏。

PR曲线：通过一系列阈值将显著图二值化后通过与真值图的对比生成PR曲线。

F-measure：对于精确率(Pre)和查全率(Rec)而言，两者成负相关关系，为了平衡两者之间的影响，采用F-measure评估实验效果，其计算公式为

$$F_\beta = \frac{(1+\beta^2)\cdot \mathrm{Pre}\cdot \mathrm{Rec}}{\beta^2\cdot \mathrm{Pre}+\mathrm{Rec}} \tag{10}$$

同文献[20]，β^2设为0.3。

MAE：平均绝对误差(MAE)评估显著图和真值图逐像素之间的绝对误差的均值。其计算公式为

$$\mathrm{MAE} = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y)-G(x,y)|}{W\cdot H} \tag{11}$$

其中，H和W分别为图像的长度和宽度，$S(x,y)$和$G(x,y)$分别表示像素$P(x,y)$的显著值和真值。

S-measure[21]：结构相似性评估准则同时评估显著图和真值图之间的区域相似性和目标相似性，其定义为

$$S_\lambda = \lambda\cdot S_0 + (1-\lambda)\cdot S_\mathrm{r} \tag{12}$$

其中，S_0和S_r分别表示目标相似性和区域相似性，λ表示平衡参数，实验中取0.5。

E-measure[22]：E-measure测量图像级别的统计信息和局部像素的匹配信息。

为了与其他方法的公平对比，所有的评估标准采用文献[23]提供的代码测试。

3.3 实验细节

本文使用python和caffe工具箱[24]进行实验，其使用的机器配置为GTX Titan-x GPUs (12 GB)。实验训练冲量、学习率和权值衰减率和最小批大小分别设置为0.99, 1e-10, 0.0005和1。本文网络结构基于两个预训练好的VGG16Net网络[9]，通过以其为初始权值并且微调，模型训练迭代10个周期，共计16万次，约耗时8 h，得到最终的网络模型。

3.4 实验对比

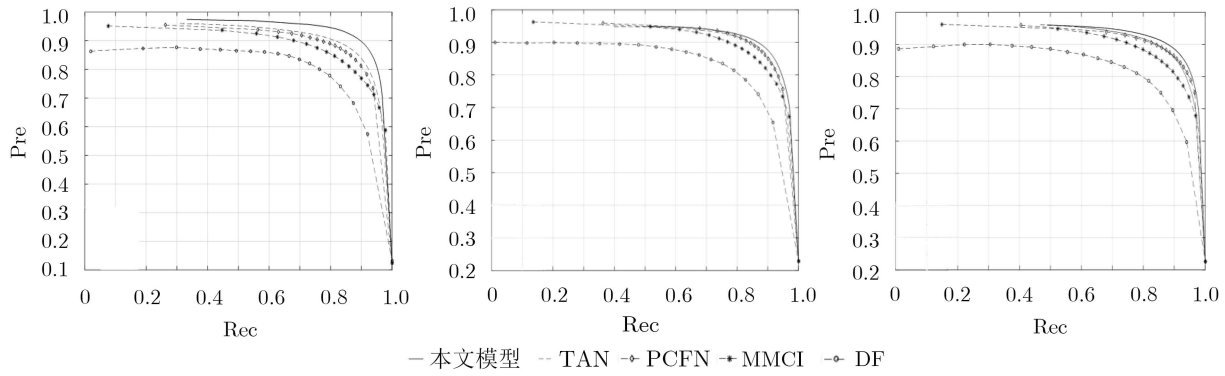
本文模型与其他的TAN[25], PCFN[8], MMCI[5], DF[4]模型在以上评估标准下进行对比，其显著图由相应论文提供或者由其提供的代码生成。

本文模型与4种具有代表性的基于深度学习的模型在PR曲线上对比，如图4所示。从图中可以看出，本文模型相对于这4种模型而言有了明显提升，并且在其他评估标准上普遍优于这4种模型。表1为本文模型实验结果在3个数据集上依据F-measure, MAE, S-measure, E-measure4个评价准则与其他模型对比，F, S, E值越高越好，MAE值越小越好，由表1中结果可以看出本文模型普遍优于其他模型，说明本文模型具有更高的泛化性。

此外，图5为本文与上述图4选取的4种模型的可视化对比结果，可以看出本文模型可以产生更好的边界信息和突出显著目标。

3.5 侧边监督模块实验对比

对第2.1节双流侧边监督模块进行实验对比，如表2所示，NDS(No Deep Supervised)表示对侧



(a) NLPR1000数据集　(b) NJU2000数据集　(c) STEREO数据集

图 4 与4种模型的PR曲线对比

边输出监督时没有使用上述监督模块，只使用1次卷积使侧边特征通道数变成1来进行监督。由表2实验结果可以看出本文监督模块可以在监督过程中更好地保留特征信息，并且生成更好的层次特征表征来满足各阶段特征需求。

3.6 多模态特征融合实验对比

对第2.2节多模态特征融合进行实验对比，如表3所示，BN表示不使用改进的多尺度卷积模块，仅采用普通卷积方式所得到的结果，从对比结果发现，在F-measure, MAE, S-measure, E-measure 4个评价准则下，多尺度卷积对显著计算结果有着重要的作用。

3.7 低维Depth特征实验对比

在实验中发现位于低维的Depth特征并不好，并且会影响最终的结果，如图6、图7所示，Depth1到Depth5表示Depth流网络各阶段对应的显著图，RGB1到RGB5表示RGB流网络各阶段对应的显著图，Conv1到Conv5表示多模态特征结合生成的显著图。从图6中可以看出由于Depth流网络的低维Depth特征，Depth1和Depth2影响导致低层输出即最终显著预测结果会出现噪声。因此为了消除低维Depth特征带来的负面效果，本文选择在低层特征融合中不加入低维Depth特征，即不加入Depth1和Depth2。

如图7所示，在没有低维Depth特征的影响下本文采用RGB1作为最终结果，从可视化对比可以看出效果有了显著提升，消除了噪声的负面影响。表4为具体的数据对比，尽管在NLPR1000数据集

表 1 在F-measure, MAE, S-measure, E-measure上与其他模型的对比

算法	NLPR1000				NJU2000				STEREO			
	F	MAE	S	E	F	MAE	S	E	F	MAE	S	E
TAN	0.7956	0.0410	0.8861	0.9161	0.8442	0.0605	0.8785	0.8932	0.8489	0.0591	0.8775	0.9108
PCFN	0.7948	0.0437	0.8736	0.9163	0.8440	0.0591	0.8770	**0.8966**	0.8450	0.0606	0.8800	0.9054
MMCI	0.7299	0.0591	0.8557	0.8717	0.8122	0.0790	0.8581	0.8775	0.8120	0.0796	0.8599	0.8896
DF	0.7348	0.0891	0.7909	0.8600	0.7703	0.1406	0.7596	0.8383	0.7650	0.1395	0.7664	0.8438
本文模型	**0.8629**	**0.0318**	**0.9117**	**0.9464**	**0.8578**	**0.0541**	**0.8852**	0.8956	**0.8622**	**0.0519**	**0.8894**	**0.9130**

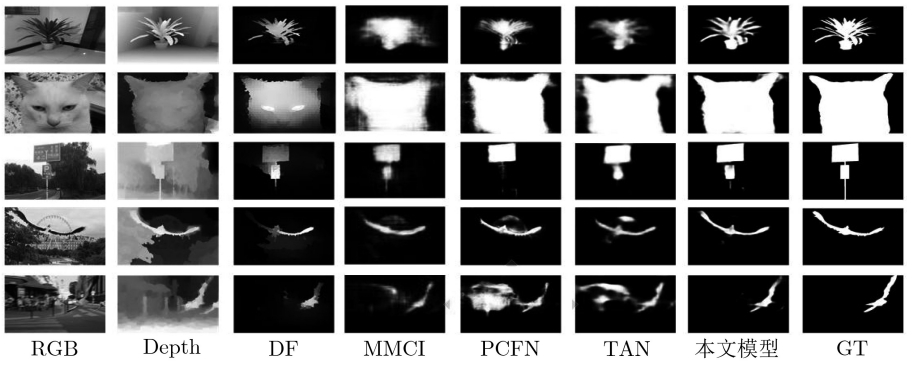


图 5 与4种模型的可视化对比

表 2 双流侧边监督模块有效性实验对比结果

算法	NLPR1000				NJU2000				STEREO			
	F	MAE	S	E	F	MAE	S	E	F	MAE	S	E
NDS	0.8358	0.0340	0.9085	0.9336	0.8502	0.0568	0.8848	0.8902	0.8524	0.0552	0.8879	0.9066
本文模型(DS)	**0.8629**	**0.0318**	**0.9117**	**0.9464**	**0.8578**	**0.0541**	**0.8852**	**0.8956**	**0.8622**	**0.0519**	**0.8894**	**0.9130**

表 3 多尺度模块有效性实验对比结果

算法	NLPR1000				NJU2000				STEREO			
	F	MAE	S	E	F	MAE	S	E	F	MAE	S	E
BN	0.8488	0.0340	0.9059	0.9398	0.8504	0.0566	0.8814	0.8928	0.8573	0.0547	0.8848	0.9093
本文模型	**0.8629**	**0.0318**	**0.9117**	**0.9464**	**0.8578**	**0.0541**	**0.8852**	**0.8956**	**0.8622**	**0.0519**	**0.8894**	**0.9130**

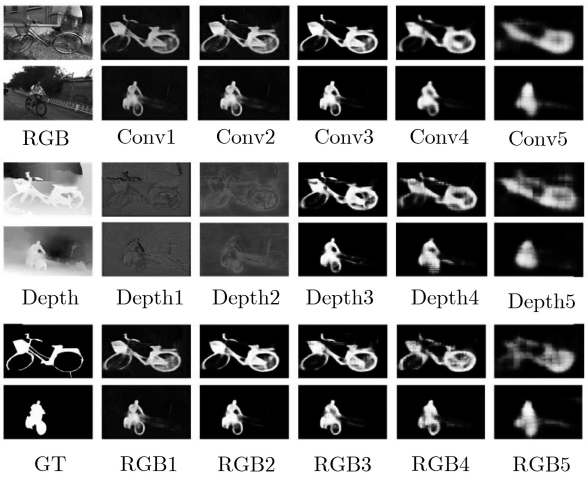


图 6 DY可视化

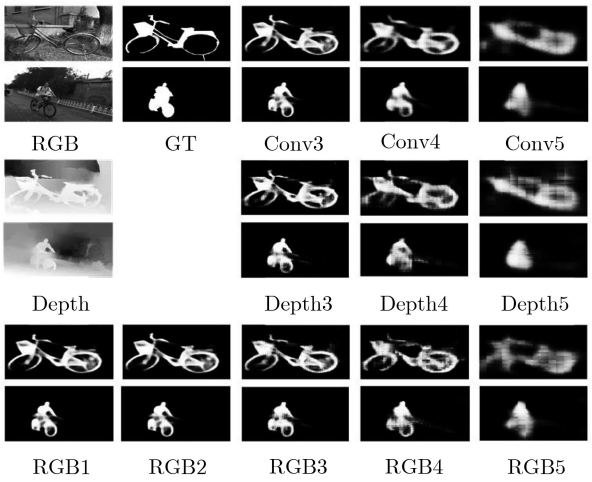


图 7 本文模型可视化

表 4　低维Depth特征实验对比结果

算法	NLPR1000				NJU2000				STEREO			
	F	MAE	S	E	F	MAE	S	E	F	MAE	S	E
DY	**0.8715**	0.1087	0.8187	**0.9479**	0.8250	0.1310	0.8414	0.8785	0.8355	0.1277	0.8541	0.8984
本文模型	0.8629	**0.0318**	**0.9117**	0.9464	**0.8578**	**0.0541**	**0.8852**	**0.8956**	**0.8622**	**0.0519**	**0.8894**	**0.9130**

上，去掉低维Depth特征后各项评价指标提升并不明显，但是综合3个数据集，最终选择没有低维Depth特征的网络作为最终的网络。

4　结束语

本文提出了一种基于全卷积神经网络模型的RGB-D显著性检测方法，包含两个模块来辅助网络以高层指导低层、从全局到局部，由深到浅得到更好的显著性预测结果。其中侧边监督模块促进网络收敛速度并且生成更好的层次表征以满足各阶段特征需求，多模态特征融合模块获取网络高层目标多尺度纹理信息，并将鲁棒的高层次特征表征进行跨模态信息互补与融合，对网络低层特征提出与高层特征不相同的融合方式，且通过文献及实验发现由于低层Depth信息不好导致最终预测结果出现噪声，故而在最终特征融合中删除低两层的Depth流特征。在3个广泛使用数据集上的实验结果表明，本文方法的实验效果普遍优于目前的主流算法，具有较强鲁棒性。未来研究中可以考虑如何使用Depth信息使其更有效地辅助RGB信息得到更好的RGB-D显著性预测结果。

参 考 文 献

[1] SHAO Ling and BRADY M. Specific object retrieval based on salient regions[J]. *Pattern Recognition*, 2006, 39(10): 1932–1948. doi: 10.1016/j.patcog.2006.04.010.

[2] GUO Chenlei and ZHANG Liming. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression[J]. *IEEE Transactions on Image Processing*, 2010, 19(1): 185–198. doi: 10.1109/TIP.2009.2030969.

[3] MAHADEVAN V and VASCONCELOS N. Biologically inspired object tracking using center-surround saliency mechanisms[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(3): 541–554. doi: 10.1109/TPAMI.2012.98.

[4] QU Liangqiong, HE Shengfeng, ZHANG Jiawei, *et al*. RGBD salient object detection via deep fusion[J]. *IEEE Transactions on Image Processing*, 2017, 26(5): 2274–2285. doi: 10.1109/TIP.2017.2682981.

[5] CHEN Hao, LI Youfu, and SU Dan. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection[J]. *Pattern Recognition*, 2019, 86: 376–385. doi: 10.1016/j.patcog.2018.08.007.

[6] HAN Junwei, CHEN Hao, LIU Nian, *et al*. CNNs-Based RGB-D saliency detection via cross-view transfer and multiview fusion[J]. *IEEE Transactions on Cybernetics*, 2018, 48(11): 3171–3183. doi: 10.1109/TCYB.2017.2761775.

[7] CHEN Hao, LI Youfu, and SU Dan. RGB-D saliency detection by multi-stream late fusion network[C]. The 11th International Conference on Computer Vision Systems, Shenzhen, China, 2017: 459-468. doi: 10.1007/978-3-319-68345-4_41.

[8] CHEN Hao and LI Youfu. Progressively complementarity-

aware fusion network for RGB-D salient object detection[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 3051–3060.

[9] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. 2015 International Conference on Learning Representations, San Diego, USA, 2015: 1150–1210.

[10] LEE C Y, XIE Saining, GALLAGHER P, *et al*. Deeply-supervised nets[C]. The 18th International Conference on Artificial Intelligence and Statistics, San Diego, USA, 2015: 562–570.

[11] XIE Saining and TU Zhuowen. Holistically-nested edge detection[J]. *International Journal of Computer Vision*, 2017, 125(1/3): 3–18. doi: 10.1007/s11263-017-1004-z.

[12] HOU Qibin, CHENG Mingming, HU Xiaowei, *et al*. Deeply supervised salient object detection with short connections[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(4): 815–828. doi: 10.1109/TPAMI.2018.2815688.

[13] DU Dapeng, XU Xiangyang, REN Tongwei, *et al*. Depth images could tell us more: Enhancing depth discriminability for RGB-D scene recognition[C]. 2018 IEEE International Conference on Multimedia and Expo, San Diego, USA, 2018: 1–6. doi: 10.1109/ICME.2018.8486573.

[14] SONG Xinhang, HERRANZ L, and JIANG Shuqiang. Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs[C]. The 31st AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017: 4271–4277.

[15] LIU Nian and HAN Junwei. DHSnet: Deep hierarchical saliency network for salient object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 678–686. doi: 10.1109/CVPR.2016.80.

[16] KIM H J, DUNN E, and FRAHM J M. Learned contextual feature reweighting for image geo-localization[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 3251–3260. doi: 10.1109/CVPR.2017.346.

[17] PENG Houwen, LI Bing, XIONG Weihua, *et al*. RGBD salient object detection: A benchmark and algorithms[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 92–109. doi: 10.1007/978-3-319-10578-9_7.

[18] JU Ran, GE Ling, GENG Wenjing, *et al*. Depth saliency based on anisotropic center-surround difference[C]. 2014 IEEE International Conference on Image Processing, Paris, France, 2014: 1115–1119. doi: 10.1109/ICIP.2014.7025222.

[19] NIU Yuzhen, GENG Yujie, LI Xueqing, *et al*. Leveraging stereopsis for saliency analysis[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 454–461. doi: 10.1109/CVPR.2012.6247708.

[20] MARTIN D R, FOWLKES C C, and MALIK J. Learning to detect natural image boundaries using local brightness, color, and texture cues[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(5): 530–549. doi: 10.1109/TPAMI.2004.1273918.

[21] FAN Dengping, CHENG Mingming, LIU Yun, *et al*. Structure-measure: A new way to evaluate foreground maps[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 4558–4567.

[22] FAN Dengping, GONG Cheng, CAO Yang, *et al*. Enhanced-alignment measure for binary foreground map evaluation[C]. The 27th International Joint Conference on Artificial Intelligence, Stockholm, 2018: 698–704.

[23] FAN Dengping, CHENG Mingming, LIU Jiangjiang, *et al*. Salient objects in clutter: Bringing salient object detection to the foreground[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 186–202.

[24] JIA Yangqing, SHELHAMER E, DONAHUE J, *et al*. Caffe: Convolutional architecture for fast feature embedding[C]. The 22nd ACM International Conference on Multimedia, Orlando, USA, 2014: 675–678. doi: 10.1145/2647868.2654889.

[25] CHEN Hao and LI Youfu. Three-stream attention-aware network for RGB-D salient object detection[J]. *IEEE Transactions on Image Processing*, 2019, 28(6): 2825–2835. doi: 10.1109/TIP.2019.2891104.

刘政怡：女，1978年生，副教授，研究方向为计算机视觉.

段群涛：女，1993年生，硕士生，研究方向为图像显著性检测.

石 松：男，1993年生，硕士生，研究方向为图像显著性检测.

赵 鹏：女，1976年生，副教授，研究方向为智能信息处理、机器学习.