# 轻量级显著性目标检测 MobileSal 网络研究报告

凌秋远

**摘要**

卷积神经网络 (CNN) 最近被广泛应用于各个领域，应用的效果非常好，但神经网络极高的计算成本阻碍了在 RGB-D 显著目标检测 (SOD) 领域的应用。MobileSal 神经网络专注于使用轻量级网络进行深度特征提取的高效 RGB-D 显著目标检测。然而，轻量级网络在特征表示方面不如“笨重”的网络强大。因此 MobileSal 还利用了彩色图像的深度信息，提出了一种隐式深度恢复 (IDR) 技术 (仅在训练时使用) 来增强轻量级网络对 RGB-D SOD 的特征表示能力。此外，MobileSal 提出了紧凑金字塔细化 (CPR) 用于高效的多层次特征融合，以导出具有清晰边界的显著对象。最终，MobileSal 在五个具有挑战性的 RGB-D SOD 数据集上表现优异，速度更快，参数更少。

**关键词**：MobileSal；SOD(salient object detection)；轻量化网络

## 1 引言

显著性目标检测 (SOD) 的目的是定位和分割自然图像中最引人注目的物体，它是图像理解中的一个基本问题，是许多计算机视觉任务 (如视觉跟踪、内容感知图像编辑和弱监督式学习) 的基础。以前的 SOD 方法主要针对 RGB 图像[1]、[2]开发，这些图像通常受到无法区分的前景和背景纹理的阻碍。为此，研究人员将容易获取的深度信息作为 RGB 图像信息的重要补充，即 RGB-D SOD。

虽然卷积神经网络 (CNN) 在 RGB-D SOD 上取得了很好的效果[3],[4]，但它们的高精度往往以高计算成本为代价，这种情况阻止了其在资源极度受限的移动设备上的部署与应用。因此，设计出兼顾精度和轻量化的 RGB-D SOD 网络至关重要。为了实现这一目标，MobileSal 需要采用轻量级骨干网络，如 MobileNets[5]、[6]和 ShuffleNets[7]、[8]进行深度特征提取，而不是通常使用的笨重骨干网络，例如 VGG[9]和 ResNets[10]。问题是，轻量级网络在特征表示学习方面通常不如笨重的网络强大，这个问题将影响轻量级网络的 RGB-D SOD 性能。

为了弥补轻量级网络在特征表示学习方面的劣势，MobileSal 提出了一种隐式深度恢复 (IDR) 技术来加强轻量级 backbone 的特征表示学习，以确保网络的准确性。具体来说，MobileSal 实现了从高层特征恢复深度图的模型，通过该模型使得轻量级 backbone 的特征学习变得更加强大，并对深度流进行了重要的监督。MobileSal 还提出了另外两个方法来确保高效率: i) 只在最粗糙的层次上进行 RGB 和深度信息融合，因为较小的特征分辨率 (1/32) 对于降低计算成本是关键的; ii) 提出了一个紧凑的金字塔细化 (CPR) 模块来有效地融合多尺度深度特征，以使得 SOD 具有清晰的边界。

## 2 相关工作

### 2.1 显著性目标检测

受益于近年来深度神经网络的快速发展，基于神经网络的 RGB 图像 SOD 方法与传统方法相比，取得了实质性进展。在这个方向上，研究人员关注于设计各种有效的策略来融合多层次 CNN 层生成

的多尺度特征[11]。尽管有许多成功的例子，但 RGB SOD 始终受到难以区分的前景和背景纹理的阻碍，这其实可以通过结合深度信息（即 RGB-D SOD）来解决。

### 2.2 RGB-D 显著性目标检测

像早期的 SOD 方法一样，传统的 RGB-D SOD 工作从 RGB 和深度图中人工提取特征，并将它们融合在一起[12]。最近，基于深度学习的 RGB-D SOD 得到了快速发展[13]。就 RGB 和深度信息的融合策略而言，RGB-D SOD 可大致分为晚期融合、早期融合和多尺度融合。晚期融合出现在特征提取结束时，仅预测融合特征的结果；早期融合直接连接输入 RGB 图像和深度图，然后从 RGB-D 输入导出显著性图；多尺度融合首先分别提取 RGB 和深度特征，然后再聚合 RGB-D 特征。尽管早期融合策略更有效，但多尺度融合更准确。为了确保高效率，MobileSal 仅在较小分辨率下最粗糙的级别融合 RGB 和深度特征，然后应用 IDR 以无计算的方式加强轻量网络的特征表示学习。

### 2.3 高效的骨干网络

越来越多的移动设备如：自动驾驶车辆、机器人和智能手机等，只有有限的计算资源，因此传统的笨重网络，如 VGG[9]和 ResNets[10]，不适合这些平台。为此，提出了一些用于图像分类的轻量网络，如 MobileNets[6]、ShuffleNets[8]、MnasNet[14]等。这些高效的网络具有低计算成本的特点，因此对于移动平台更适用。在 MobileSal 中首次通过采用 MobileNetV2[6]作为深度特征提取的骨干网络，实现了高效 RGB-D SOD。

## 3 本文方法

### 3.1 概述

在 RGB 流中，MobileSal 使用了 MobileNetV2[6]作为骨干网络，且为了适应 SOD 任务，删除了全局平均池层和最后一个完全连接的层。对于 RGB 流，每个阶段之后都是步幅为 2 的卷积层，因此特征图在每个阶段之后被下采样为一半分辨率。五个阶段的输出特征图表示为 C1、C2、C3、C4、C5，步幅分别为 $2, 2^2, 2^3, 2^4, 2^5$.

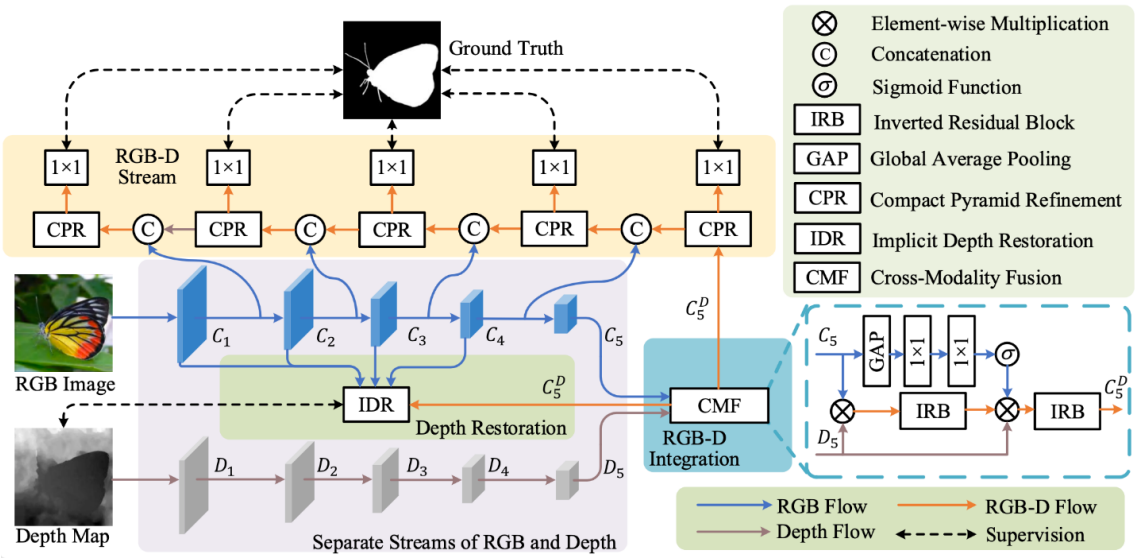


图 1: MobileSal 整体网络结构

在 Depth 流中，与 RGB 流类似，Depth 流也有五个具有相同步幅的阶段。由于深度图包含的语义信息比相应的 RGB 图像少，因此 MobileSal 构建了一个具有比 RGB 流更少卷积块的轻量级深度网络，

每个阶段只有两个倒置残差块 (IRB)[6]，这种设计降低了计算复杂度。在每个 IRB 中，首先通过 1×1 卷积将特征图沿通道扩展 M 倍，然后使用相同数量的输入和输出通道进行深度可分离的 3×3 卷积 [16]。然后，通过 1×1 卷积将特征通道压缩到 1/M。深度流的五个阶段的输出特征图表示为 D1、D2、D3、D4、D5，其中前四个阶段分别具有 16、32、64、96 个通道。D5 和 C5 具有相同数量的通道和相同的步幅。

如图 1所示，利用 RGB 和 Depth 流的输出，首先融合 RGB 特征 C5 和深度特征 D5，以生成 RGB-D 特征 $C_5^D$。IDR 模块从 C1、C2、C3、C4、$C_5^D$ 恢复深度图，该深度图由输入深度图监督以加强特征表示学习。对于显著性预测，还设计了一个以 CPR 模块为基本单元的轻量级解码器，底层解码器的输出是最终预测的显著性图。

### 3.2 跨模融合模块 CMF

深度图揭示了彩色图像的空间线索，有助于区分前景和背景，尤其是对于具有复杂纹理的场景。CMF(Cross-Modal Fusion) 模块只在最粗级别融合，仅融合 RGB 特征图 C5 和深度特征图 D5。语义信息主要存在于 RGB 图像中，深度图传达了深度平滑区域的优先级，通过乘法来增强 RGB 语义特征，这可以看作是一种强大的正则化。

首先将 RGB 和深度特征与上述 IRB 相结合，以得出过渡的 RGB-D 特征图 $\tau$，其可以表示为 $\tau = IRB(C_5 \otimes D_5)$, 其中 $\otimes$ 为元素乘法运算符；将全局平均池 (GAP) 层应用于 C5 以获得特征向量，然后两个全连接层以计算 RGB 注意力向量 v, 即

$$\mathbf{v} = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(\mathrm{GAP}(\mathcal{C}_5))))), \tag{1}$$

其中 FC 和 ReLU 分别表示全连接层和 ReLU 层。$FC_1$ 和 $FC_2$ 的输出通道数量与输入相同。$\sigma$ 表示标准 sigmoid 函数。计算 $\tau$ 和 v 后，将 v、$\tau$ 和 $D_5$ 的乘积输入 IRB，得到：

$$\mathcal{C}_5^D = \mathrm{IRB}(\mathbf{v} \otimes \mathcal{T} \otimes \mathcal{D}_5), \tag{2}$$

其中 $C_5^D$ 代表 CMF 模块的输出特征图。在融合 RGB 和深度特征之后，我们可以导出主干特征，包括 RGB 特征 C1、C2、C3、C4 和融合的 RGB-D 特征 $C_5^D$。

### 3.3 隐式深度恢复 IDR

通常来说，一个物体或一个连通的区域具有相似的深度，因此可以将深度图作为一个额外的监督源来指导表征学习，从而帮助轻量化网络抑制物体或连通物体区域内的纹理变化，突出的物体和背景之间的差异。隐式深度恢复 (IDR) 模块是基于这一思想设计的。

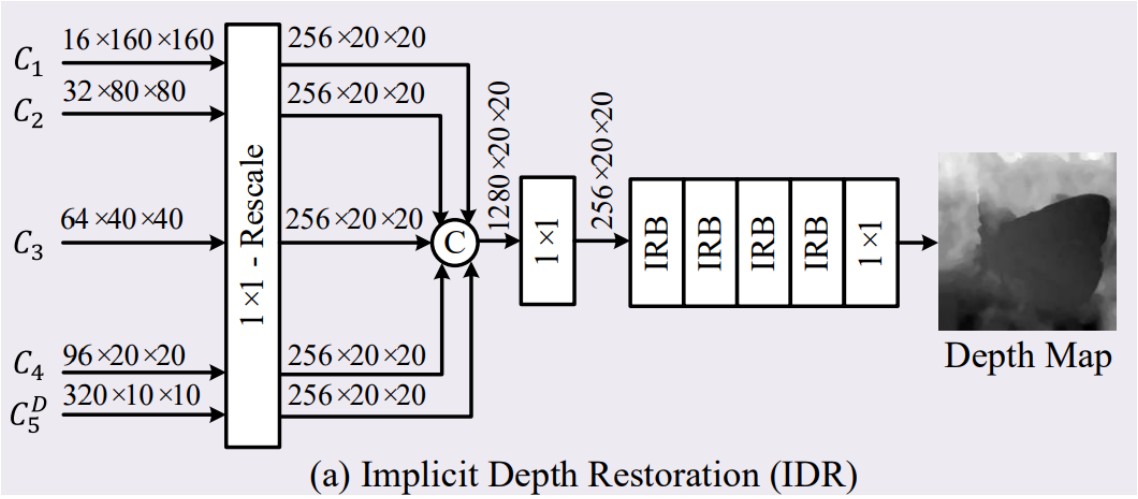


图 2: IDR 模块

IDR 使用 C1、C2、C3、C4、$C_5^D$ 进行辅助监控。如图 2所示，首先用 1×1 卷积来将 C1、C2、C3、C4、$C_5^D$ 压缩到相同的通道 (256)。然后，将生成的特征图调整为与 C4 相同的大小将它们串联起来。1×1 卷积将级联特征图从 1280 个通道更改为 256 个通道，以节省计算成本。接下来，用四个 IRB 融合多层次特征，获得强大的多尺度特征。最后 1×1 卷积将融合的特征图转换为单通道。通过标准的 sigmoid 函数和双线性上采样，获得与输入大小相同的恢复深度图。

### 3.4 紧凑金字塔细化 CPR

骨干网络中的高级特征包含语义抽象特征，而低级特征传递细粒度细节。对于准确的 SOD，必须充分利用高层和低层特征。

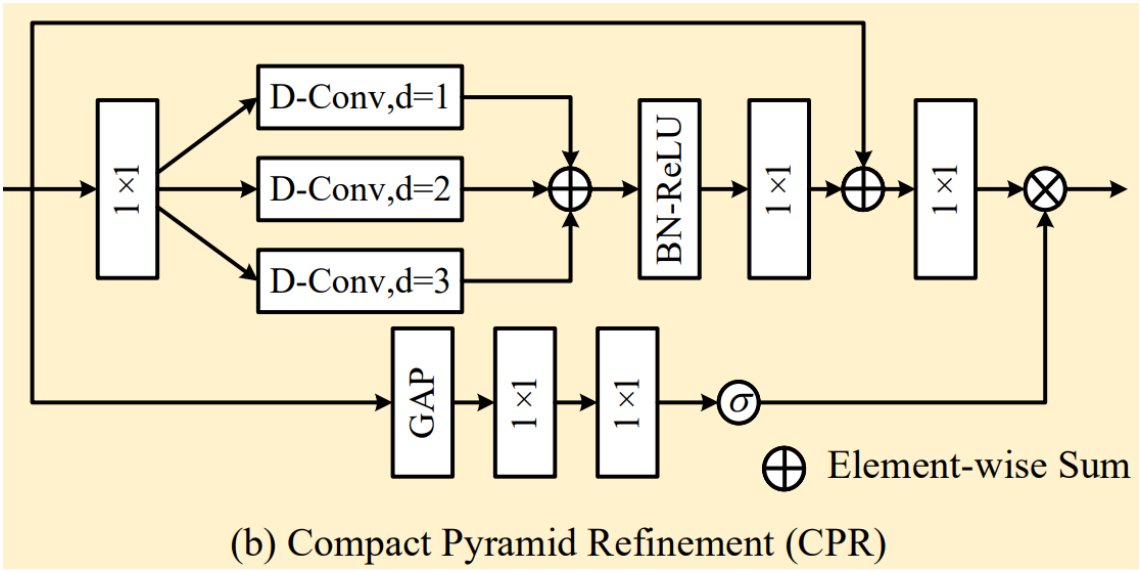


图 3: CPR 模块

CPR 采用了一种轻量级的多尺度学习策略来增强融合。假设 CPR 模块的输入为 X。如图 3所示，CPR 首先用 1×1 卷积将通道数量扩展 M 倍，然后并行连接三个 3×3 深度可分离卷积 (膨胀率为 1，2，

3)，用于多尺度融合。这可以表述为

$$\begin{aligned}\mathcal{X}_1 &= \text{Conv}_{1\times1}(\mathcal{X}),\\ \mathcal{X}_2^{d_1} &= \text{Conv}_{3\times3}^{d_1}(\mathcal{X}_1),\\ \mathcal{X}_2^{d_2} &= \text{Conv}_{3\times3}^{d_2}(\mathcal{X}_1),\\ \mathcal{X}_2^{d_3} &= \text{Conv}_{3\times3}^{d_3}(\mathcal{X}_1),\\ \mathcal{X}_2 &= \text{ReLU}(\text{BN}(\mathcal{X}_2^{d_1}+\mathcal{X}_2^{d_2}+\mathcal{X}_2^{d_3})),\end{aligned} \tag{3}$$

其中，$d_1$、$d_2$ 和 $d_3$ 分别是扩张率。再使用 1×1 卷积将通道压缩到与输入相同的数量。

$$\mathcal{X}_3 = \text{Conv}_{1\times1}(\mathcal{X}_2) + \mathcal{X}, \tag{4}$$

使用残余连接以获得更好的优化。最终重新校准融合的特征，有

$$\mathcal{Y} = \mathbf{v}' \otimes \text{Conv}_{1\times1}(\mathcal{X}_3). \tag{5}$$

其中，v′ 为注意力向量。

在每个解码器阶段，来自顶部解码器和相应的编码器阶段的两个特征映射首先使用 1×1 卷积，然后将结果连接起来，接着用一个 CPR 模块进行特征融合。通过这种方式，轻量级解码器实现从上到下聚合了多个级别的特性。

## 4 复现细节

### 4.1 与已有开源代码对比

原论文中用到的 backbone 是 MobileNetV2 网络，而 MobileNetV3[15]是 MobileNetV2 的更新版本，保留了 V1[5]的深度可分离卷积和 V2 的线性瓶颈倒残差结构思想，重新设计了耗时层结构，减少第一个卷积层的卷积核个数，并引入基于 squeeze and excitation 结构的轻量级注意力模型 (SE)，使用了一种新的激活函数 $h-swish(x)$ 。

在修改了 backbone 后，提取了通道数为：$[24, 40, 80, 160, 960]$ 这 5 层特征，分别对应 $C1, C2, C3, C4, C5$。并更改、对齐了原论文 CMF、IDR、CPR 模块中的通道数。

| Input | Operator | exp size | $\#out$ | SE | NL | $s$ |
|---|---|---|---|---|---|---|
| $224^2 \times 3$ | conv2d | - | 16 | - | HS | 2 |
| $112^2 \times 16$ | bneck, 3x3 | 16 | 16 | - | RE | 1 |
| $112^2 \times 16$ | bneck, 3x3 | 64 | 24 | - | RE | 2 |
| $56^2 \times 24$ | bneck, 3x3 | 72 | 24 | - | RE | 1 |
| $56^2 \times 24$ | bneck, 5x5 | 72 | 40 | ✓ | RE | 2 |
| $28^2 \times 40$ | bneck, 5x5 | 120 | 40 | ✓ | RE | 1 |
| $28^2 \times 40$ | bneck, 5x5 | 120 | 40 | ✓ | RE | 1 |
| $28^2 \times 40$ | bneck, 3x3 | 240 | 80 | - | HS | 2 |
| $14^2 \times 80$ | bneck, 3x3 | 200 | 80 | - | HS | 1 |
| $14^2 \times 80$ | bneck, 3x3 | 184 | 80 | - | HS | 1 |
| $14^2 \times 80$ | bneck, 3x3 | 184 | 80 | - | HS | 1 |
| $14^2 \times 80$ | bneck, 3x3 | 480 | 112 | ✓ | HS | 1 |
| $14^2 \times 112$ | bneck, 3x3 | 672 | 112 | ✓ | HS | 1 |
| $14^2 \times 112$ | bneck, 5x5 | 672 | 160 | ✓ | HS | 2 |
| $7^2 \times 160$ | bneck, 5x5 | 960 | 160 | ✓ | HS | 1 |
| $7^2 \times 160$ | bneck, 5x5 | 960 | 160 | ✓ | HS | 1 |
| $7^2 \times 160$ | conv2d, 1x1 | - | 960 | - | HS | 1 |
| $7^2 \times 960$ | pool, 7x7 | - | - | - | - | 1 |
| $1^2 \times 960$ | conv2d 1x1, NBN | - | 1280 | - | HS | 1 |
| $1^2 \times 1280$ | conv2d 1x1, NBN | - | k | - | - | 1 |

图 4: MobileNetV3 网络结构。SE: Squeeze-And-Excite, NL: nonlinearity, HS: h-swish, RE: ReLU

### 4.2 实验环境

最终，在 pytorch 中复现了更改后的 MobileSal 网络。其中，将 RGB 和深度图像的大小调整为 320×320，并使用水平翻转和随机裁剪作为数据增强。还应用了多尺度训练，即每个图像的大小调整为 $[256, 288, 320]$ 训练，但保持测试图像的大小不变。复现时使用了单个 GTX 1050Ti 图形处理器进行训练和测试，初始学习率 lr 设置为 0.0001，$batchSize$ 大小设置为 4，并采用 $Adam$ 优化器用于优化网络。

复现时使用了 5 个具有代表性的公开数据集，分别为：NJU2K，NLPR，STERE，SSD 和 SIP。它们分别包含 1985、1000、1000、80、927 张图片。其中，使用了 1500 张 NJU2K 图像和 700 张 NLPR 图像进行训练，另外 485 张 NJU2K 图像和 300 张 NLPR 图像进行测试。

## 5 实验结果分析

用训练好的网络在 5 个具有代表性的数据集上测试，结果如图 5所示。其中，每一行代表不同的数据集（包含 NJU2K，NLPR，STERE，SSD 和 SIP）；第一列是 RGB 图像，第二列是对应的深度图像，第三列为标签，第四列为原论文中用的 MobileNetV2 的结果，第五列为修改 backbone 为 MobileNetV3 后结果。对比结果可以看出，除了在 SSD 数据集上显著性目标的检测效果不太理想，在其它数据集上的效果均不错，基本吻合。

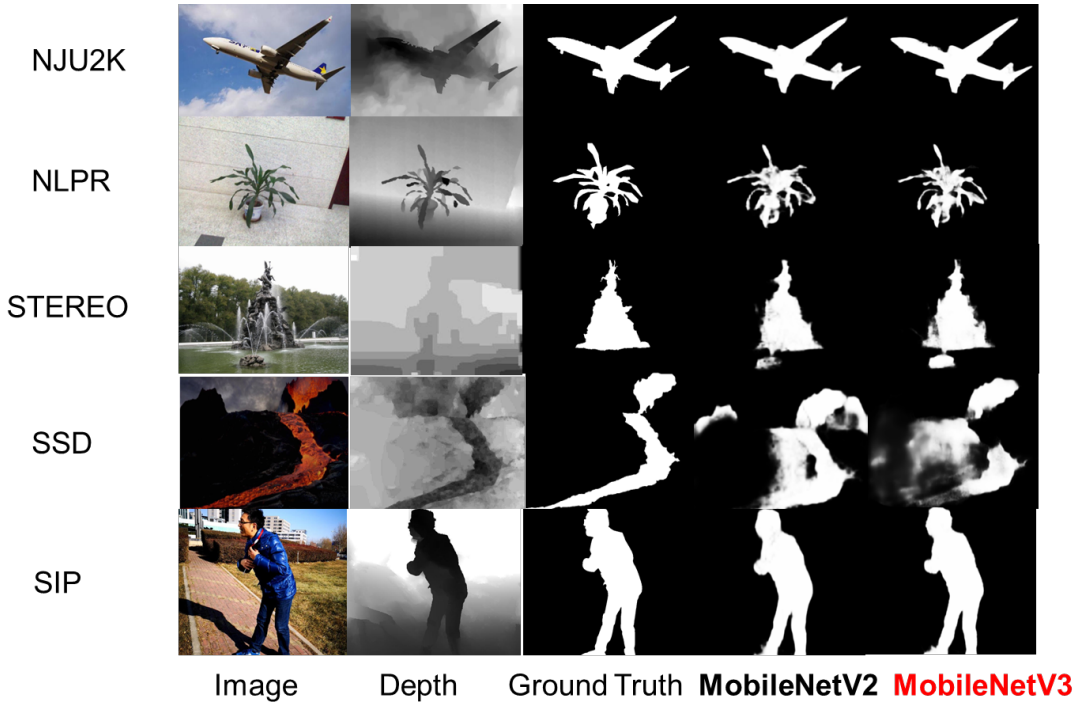


图 5: 最终复现结果对比

最终的测试精度表格如图 6所示 (图中红色数字表示更优)。由表格可以看出，在单张 GTX 1050Ti GPU 上测试的帧率达到了 12 FPS 左右；总体上，在这 5 个数据集上测试的精度 $F_\beta$ 基本达到了 90% 左右；更改了 backbone 为 MobileNetV3 后，对比 V2 并没有明显的提升，无论在精度还是速度上均与原论文的结果接近。

| | | 单尺度训练(Single-scale Training) 320*320 | | 多尺度训练 (Multi-scale Training) [256, 288, 320] | |
|---|---|---|---|---|---|
| 方法 | | MobileSal mobilenetV2 | MobileSal mobilenetV3_large | MobileSal mobilenetV2 | MobileSal mobilenetV3_large |
| Speed (fps) | | 12.17 | 11.68 | 11.91 | 11.44 |
| Params (M) | | 6.5 | 40.1 | 6.5 | 40.1 |
| NJU2K | F_beta | 0.9037 | 0.9061 | 0.9140 | 0.9072 |
| | MAE | 0.0452 | 0.0446 | 0.0408 | 0.0475 |
| NLPR | F_beta | 0.9132 | 0.9122 | 0.9172 | 0.9181 |
| | MAE | 0.0255 | 0.0245 | 0.0247 | 0.0276 |
| STEREO | F_beta | 0.9050 | 0.9008 | 0.9067 | 0.9074 |
| | MAE | 0.0427 | 0.0429 | 0.0404 | 0.0402 |
| SSD | F_beta | 0.8405 | 0.8481 | 0.8624 | 0.8616 |
| | MAE | 0.0560 | 0.0528 | 0.0521 | 0.0508 |
| SIP | F_beta | 0.8959 | 0.8895 | 0.8986 | 0.8965 |
| | MAE | 0.0543 | 0.0559 | 0.0529 | 0.0540 |

图 6: 测试精度汇总

## 6 总结与展望

MobileSal 是一个非常高效的 RGB-D SOD 网络，不仅精度高，速度还快，非常适合在资源受限的边缘设备上部署。在复现的过程中，虽然把骨干网络 MobileNetV2 更新到了 MobileNetV3，但并没有对原网络有明显改善，希望日后能找到改进的新方向。

## 参考文献

[1] ZHANG P, WANG D, LU H, et al. Amulet: Aggregating multi-level convolutional features for salient object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 202-211.

[2] PANG Y, ZHAO X, ZHANG L, et al. Multi-scale interactive network for salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9413-9422.

[3] ZHAO X, ZHANG L, PANG Y, et al. A single stream network for robust and real-time RGB-D salient object detection[C]//European Conference on Computer Vision. 2020: 646-662.

[4] FU K, FAN D P, JI G P, et al. Siamese network for RGB-D salient object detection and beyond[J]. IEEE transactions on pattern analysis and machine intelligence, 2021.

[5] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[6] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[7] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.

[8] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] ZHAO X, PANG Y, ZHANG L, et al. Suppress and balance: A simple gated network for salient object detection[C]//European conference on computer vision. 2020: 35-51.

[12] LANG C, NGUYEN T V, KATTI H, et al. Depth matters: Influence of depth cues on visual saliency[C]//European conference on computer vision. 2012: 101-115.

[13] ZHAO J X, CAO Y, FAN D P, et al. Contrast prior and fluid pyramid integration for RGBD salient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3927-3936.

[14] TAN M, CHEN B, PANG R, et al. Mnasnet: Platform-aware neural architecture search for mobile[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2820-2828.

[15] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.