Res2Net: 一种新的多尺度骨干架构

Shang-Hua Gao*, Ming-Ming Cheng*, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr

摘要—对于众多的计算机视觉任务来说，表达出多尺度特征是非常重要的。最近卷积神经网络（CNNs）骨干的一些进展，不断表现出来了更强的多尺度特征表达能力。这也使得其在广泛的使用领域中得到了性能的提升。不过，现存的大多数方法表达多尺度特征都是通过逐层的方式进行的。在本文中，我们提出了一种新颖的架构卷积神经网络模块的方法，我们称之为 Res2Net 。它通过在一个残差块中构筑类似残差分层的方式进行连接。我们的 Res2Net 可以在更细粒度级别表达多尺度特征，并且可以增加每层网络的感受野大小。我们提出的这个残差块可以被加入到类似 ResNet，ResNeXt，DLA 等一流的网络结构中。为上述模型加入 Res2Net 模块后，我们在如 CIFAR-100 和 ImageNet 等知名数据集上获得了比原结构更好性能。进一步的研究表明，在一些如物体检测，CAM，显著性物体检测等具有代表性的视觉任务中，我们的 Res2Net 具有比其他基础模型更好的表现。源代码和训练好的模型已经在 https://mmcheng.net/res2net/开源。

关键词 – 多尺度，深度学习。

1 引言

如图1所示，自然界的图形都是多尺度的。首先，如沙发、杯子等，他们都是出现在一张图片中的大小不同的物体。其次，一个物体所需要的上下文语境信息可能比它本身的面积占比要大。举个例子，我们需要图中的桌子作为上下文语境信息来区分桌上的黑色小物体是笔筒还是杯子。第三，感受来自不同尺度的信息有助于细分类和语义分割等任务的性能提升。因此，在很多的计算机视觉任务中，如图像分类 [28]、物体检测 [43]、注意力预测 [45]、目标追踪 [45]、动作识别 [46]、语义分割 [6]、显著性物体检测 [2], [24]、物体提取 [12], [43]、骨架提取 [66]、立体匹配 [42]、边缘检测 [37], [57] 等，设计一个优秀的可以表达多尺度特征的结构是十分重要的。

多尺度特征已经被广泛的应用于常规的特征设计 [1], [39] 和深度学习 [10], [51] 中。在众多的视觉任务中，为了使网络获得多尺度表达能力，需要特征提取器使用一个大范围的感受野来获得不同尺度的物品/部件/上下文信息等。卷积神经网络（CNNs）通过堆叠卷积层使得网络由粗到细地学习多尺度特征。卷积神经网络的这种多尺度特征提取能力可以很有效的解决众多视觉任务中的问题。进一步提升卷积神经网络性能的关键就是需要设计一个更加有效的网络结构。

在过去的几年中，部分网络骨架结构 [10], [15], [23], [25], [26], [28], [47], [51], [56], [60] 在众多的视觉任务中取得了重大的进展，并且获得了一流性能。更早的神经网络结构如 AlexNet [28] 和 VGGNet [47] 中，他们通过堆叠卷积操作使得使用数据驱动多尺度特征的学习变得可行。之后，为了提升网络的多尺度表达能力使用了很多不同的方法，如 InceptionNets [50], [51], [52] 等网络，通过在卷积层使用不同大小的卷积核，如 ResNet [23] 使用的残差模块，DenseNet [26] 使用的快捷链路和 DLA [60] 等，这些方法均能提升多尺度表达能力。卷积神经网络的骨架结构的进步表明了，神经网络正在倾向于向更加高效和有效的多尺度表达能力方面发展。

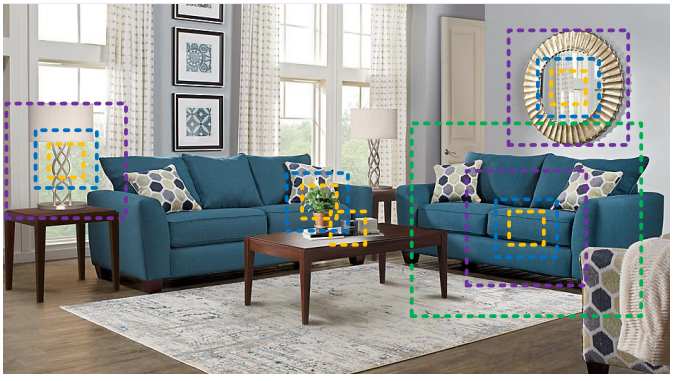


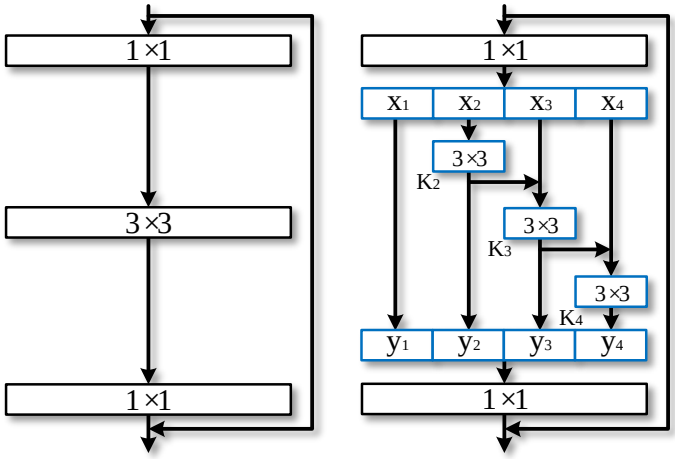
图 1. 多尺度的表达对于各种视觉任务是十分重要的，如物体边界、区域的识别和目标物体语义类别的识别等等。即使在在最简单的物体识别任务中，对于不同尺度物体 (如沙发、杯子等) 信息的感知和他们所处的上下文环境 (如我们要识别图中黑色小杯子，就需要一个“它在桌子上”的信息) 也是很重要的。

在本文中，我们提出了一种简单并且有效的多尺度处理流程。跟大多数现有的方法来提升层级的多尺度表达能力不同，我们提出的方法是在更细粒度的级别提升网络的多尺度表达能力。跟一些现有方法 [5], [9], [11] 利用具有不同分辨率的特征图来提升多尺度表达能力不同的是，我们提出的方法是在更细粒度的尺度上通过多个感受野来提升。我们通过将

*Equal contribution

- S.H. Gao, M.M. Cheng, K. Zhao, and X.Y Zhang are with the TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China.
- M.H. Yang is with UC Merced.
- P. Torr is with Oxford University.
- M.M. Cheng is the corresponding author (cmm@nankai.edu.cn).

- 本文是 IEEE TPAMI 论文 [17] 的中译版。



(a) 瓶颈层 (b) Res2Net 模块

图 2. 瓶颈层和我们提出的 Res2Net 模块对比 (假设尺度维度 $s = 4$ 时)。

原有的 n 通道 3×3 滤波器[1]替换为一系列有 w 通道的更小的滤波器组（避免损失，我们令 $n = s \times w$）。如图2，小滤波器组以类似于残差的模式被逐层连接，这样可以增加输出特征能表达的不同尺度的数量。我们将输入的特征图分为几组，每一组滤波器先从一组输入特征图中进行特征提取，然后与先前组生成的特征图和另一组输入的特征图一起被送到下一组卷积核进行处理。这个过程将一直持续到所有特征图都被处理完毕。最终，所有特征图将被拼接在一起并被送到一组 1×1 的卷积核处进行信息融合。在任意可能的将输入特征图转化为输出特征图的路径上，相等的感受野在经过 3×3 的卷积核后总会增多，最终会因组合效应生成出许多等价的的特征尺度。

我们的 Res2Net 方法扩展出来了一个**新的维度**，我们命名为*尺度维度* (scale) （Res2Net 模块中特征图被分成的组数），它将作为一个基础的参数出现在网络中，如同深度 (depth) [47]、宽度 (width)[2]和基数 (cardinality) [56] 我们将在4.4小节中演示提搞尺度维度将比提高其他维度更加有效。

请注意，我们提出的在更细粒度级别扩展一个网络的多尺度表达潜力的方法和现存的其他利用逐层操作的方法是没有冲突的。因此我们的 Res2Net 模块可以很容易就被加入到其他现存的卷积网络框架中。大量实验表明，我们的 Res2Net 模块可以很好地提升现有一流网络如 ResNet [23]、ResNeXt [56]、和 DLA [60] 等的表现。

2 相关工作

2.1 骨干网络

近几年，我们见证了拥有更强的多尺度表达能力的网络骨干结构 [15], [23], [26], [28], [47], [51], [56], [60] 在众多的视觉任务中取得了世界一流的性能水平。卷积神经网络正如被设计的那样，因为输入的信息是由细到粗的模式，它具有基本的多尺度特征表达能力。AlexNet [28] 通过顺序的堆积卷积层，在物体识别方面实现了比传统方法更大的性能突破。不过受限于其网络的深度和卷积核的大小，AlexNet 只有很小的感受野。VGGNet [47] 使用了更小的卷积核，然后增加了网络的深度。更深的网络可以获得更大感受野，有助于提取大尺度物体的特征。通过堆叠更多的网络层数往往比扩大卷积核能更有效的扩大感受野大小。正因为如此，VGGNet 比 AlexNet 使用了更少的参数却获得了更强的多尺度特征表达能力。不过 AlexNet 和 VGGNet 都是直接堆积卷积层，这会使得其每一层都有一个相对固定的感受野。

NIN [31] 通过将多层感知器作为子网络加入大型网络中，这样做增强了感受野对于其内部局部位置的辨别能力。NIN 中 1×1 的卷积层也成为了一种优秀的融合特征图信息的方法。GoogLeNet [51] 利用并行的使用不同卷积核的卷积层来增强网络的多尺度表达能力。不过由于参数量的影响，网络的性能往往受计算量约束。Inception 网络 [50], [52] 通过在 GoogLeNet 的并行路径上堆叠更多的卷积层来扩大感受野。另一方面，ResNet [23] 提出的神经网络中的短连接方式可以减缓梯度消失的现象，从而可以获得一个更深的网络结构。在特征提取过程中，短连接允许组合不同的卷积操作，这使得它会有大量等价的的特征图尺度。类似的，DenseNet [26] 中的稠密连接层使得网络可以处理大尺度的物体。DPN [10] 对于 ResNet 和 DenseNet 的组合使得其拥有 ResNet 的特征复用能力和 DenseNet 的特征提取能力。在最近的 DLA [60] 中，其以类似树的结构来组合网络的层结构。树状结构的分层使得网络也能获得更好的逐层的多尺度表达能力。

2.2 视觉任务中的多尺度表达能力

卷积神经网络中，多尺度表达能力非常重要。如在物体检测 [43]、面部分析 [4], [41]、边缘检测 [37]、语义分割 [6]、显著性物体检测 [34], [64] 和骨架检测 [66] 中，它将能有效提升模型的性能。

2.2.1 物体检测

一个有效的卷积神经网络需要能够有效的辨别出一个场景中的各种不同尺度的物体。在早期的论文如 R-CNN [18] 中，其主要依靠如 VGGNet [47] 作为骨架网络来提取多尺度特征。He 等人提出了 SPP-Net [22] 方法，它可以在骨干网络后使用空间金字塔池化来增强多尺度表达能力。后来的 Faster R-CNN [43] 提出的区域候选网络（RPN）用来生成不同尺度的边界框。基于 Faster R-CNN，FPN [32] 加入了特征图金字塔的方法来提取单个图像中不同尺度的特征信息。SSD [36] 方法利用了不同阶段的特征图来处理不同尺度的视觉信息。

2.2.2 语义分割

对于语义分割来说，通常需要卷积神经网络处理不同尺度的特征来提取上下文语境信息。Long [38] 等人最早提出了对

1. 卷积操作和滤波器两个名词同义。
2. 宽度 (width) 表示一个层中的通道数 [61].

于语义分割有效的多尺度表达的方法是全卷积网络（FCN）。在 DeepLab 中，Chen 等人 [6], [7] 提出的级联空洞卷积模块可以在保持空间分辨率的同时扩大感受野。最近的论文 PSPNet [63] 中，通过金字塔池化的方法汇总来自基于区域特征的全局上下文信息。

2.2.3 显著性物体检测

为了精准定位图像中的显著性物体所在的区域，需要模型理解大量的多尺度上下文信息以确定物体的显著性水平，并且需要小尺度的特征信息来准确定位物体的边界 [65]。更早的方法如 [3] 利用手动来维护全局对比信息 [13] 或者多尺度区域特征 [53]。Li 等人 [29] 最早提出的方法之一能够利用深度多尺度特征来进行显著性物体检测。之后，多语境深度学习 [67] 和多层次卷积特征 [62] 也被提出以提升显著性物体检测的性能。最近的论文中，Hou 等人 [24] 提出的在不同阶段加入稠密短连接来丰富每层中的多尺度信息以提升显著性物体检测的性能。

2.3 同期论文

最近也有一些论文 [5], [9], [11], [49] 来利用多尺度特征提升网络性能。Big-Little 网络 [5] 是一个由不同计算复杂度分支组成的多分支网络。OctConv [9] 将标准卷积分解为两种分辨率，以不同的频率处理特征图。MSNet [11] 通过上采样，从低分辨率网络中得到了低分辨率特征图，来在高分辨率网络中学习高频率的残差。除了部分论文中的低分辨率表达之外，HRNet [48], [49] 也在网络中引入了高分辨率表达并且反复融合多尺度特征以增强高分辨率表达。这些模型 [5], [9], [11], [48], [49] 中，都使用了池化或者上采样来使得图像的尺寸变为原来的 2^n 倍。这样可以在性能不变甚至提升基础上减小计算量。在 Res2Net 模块中，层次间类似残差的块内连接能够使得感受野不断变化，从而使得其可以在更加细粒度的水平上捕捉局部或者全局的图像特征。实验结果也表明，Res2Net 模块可以被集成在新的网络结构上以提升他们的性能。

3 Res2Net

3.1 Res2Net 模块

如图2(a) 中的模块，它是许多现代卷积神经网络如 ResNet [23], ResNeXt [56], 和 DLA [60] 的基础构建结构。在保持计算量大致不变的情况下，我们找到了一些拥有比 3×3 卷积层更强的多尺度特征提取能力的网络结构。我们将 3×3 的卷积层替换成分层小组的卷积，并且通过类似残差连接的方式将不同的卷积组逐层串起来。因为我们提出的这个模块在单个残差块中有类似**残差**的连接，所以我们将其命名为 Res2Net。

图2展示出了我们提出的 Res2Net 模块和普通瓶颈层的差别。在 1×1 的卷积层后面，我们将特征图分为 s 个子集，用 x_i 表示，其中 $i \in \{1, 2, ..., s\}$。除了子集的通道数是原来

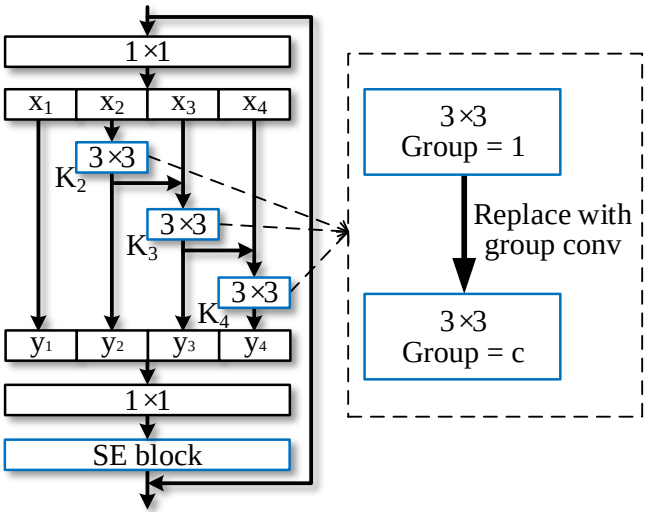


图 3. Res2Net 模块可以与基数（cardinality）维度 [56](用卷积组替换原来的卷积) 和 SE [25] 模块一起集成在模型中。

的 $1/s$ 外，每个特征图子集 x_i 都有着和原始特征图集相同的空间大小。除了 x_1，每个特征图子集 x_i 都有其对应的 3×3 卷积层，用 $\mathrm{K}_i()$ 表示。我们定义 $\mathrm{K}_i()$ 的输出为 y_i。特征图子集 x_i 和 $\mathrm{K}_{i-1}()$ 相加后被一同送入 $\mathrm{K}_i()$ 进行处理。为了减少参数并增加 s 的数量，作者忽略了对 x_1 所要进行的 3×3 卷积因此 y_i 可以被表示为：

$$\mathrm{y}_i = \begin{cases} \mathrm{x}_i & i = 1; \\ \mathrm{K}_i(\mathrm{x}_i) & i = 2; \\ \mathrm{K}_i(\mathrm{x}_i + \mathrm{y}_{i-1}) & 2 < i \leqslant s. \end{cases} \tag{1}$$

注意到每一个 3×3 卷积核 $\mathrm{K}_i()$ 可以潜移默化地接收到他之前所有特征图子集，即 $\{\mathrm{x}_j, j \leq i\}$ 的特征信息。每一次特征图子集 x_j 通过一个 3×3 的卷积核之后，输出结果就可以有一个比 x_j 更大的感受野。因为组合爆炸效应，Res2Net 模块的输出包含了不同数量的不同大小以及不同尺度的感受野与他们的不同组合。

在我们的 Res2Net 模块中，拆分出的部分是以多尺度的方式处理的，这有利于提取局部和全局的信息。为了让不同尺度的信息融合得更好，我们将拆分出的部分并联在一起然后通过一个 1×1 的卷积层进行信息融合。这种分组、合并的策略使得卷积层能够更有效地处理特征图。为了减少参数，我们忽略了第一个分组的卷积层，这也是一种特征复用的形式。

在本中文，使用参数 s 控制尺度（scale）维度。更大的 s 将能更好的学习到更丰富的不同尺寸的感受野，但是这样做将会增加计算量和内存的消耗。

3.2 在流行模型上的集成

近几年，大量的神经网络模型被提出，如 Xie 等人 [56] 引入的基数（cardinality）维度还有 Hu 等人 [25] 提出的 SE 模块。Res2Net 模块引入的尺度（scale）维度和这些方法是可以

共存的。如图3，我们可以很容易地将基数（cardinality）维度 [56] 和 SE 模块 [25] 集成到我们的 Res2Net 模块中。

3.2.1 基数（cardinality）维度

基数（cardinality）维度表示一个卷积层内的组的数量 [56]。这个维度将网络的卷积层从单分支变成了多分支，从而提升了模型的表达能力。将我们模块中 3 × 3 的卷积层替换为一组 3 × 3 的卷积组，其中用 c 表示组的数量。我们在4.2小节和4.4小节中对比了不同的的尺度（scale）维度和基数（cardinality）维度对于网络性能的影响。

3.2.2 SE 模块

SE 模块通过显式地建立通道之间的联系来自适应地调整各通道之间的特征响应 [25]。类似于 [25]，我们在 Res2Net 模块的残差连接前面加入了 SE 模块。SE 模块可以提升 Res2Net 模块的性能，这点我们将在 4.2小节和4.3小节进行演示。

3.3 模型集成

因为我们的 Res2Net 模块对于网络的基础结构没有严格的要求，并且 Res2Net 模块的这种多尺度特征能力和其他的层级的特征聚合模型可以共存，所以它就可以很容易地被集成在如 ResNet [23]、ResNeXt [56]、DLA [60] 和 Big-Little 网络 [5] 等一流的网络结构中。上述模型集成模块之后分别对应 Res2Net、Res2NeXt、Res2Net-DLA 和 bLRes2Net-50。

我们提出的尺度（scale）维度和之前已经被提出的基数（cardinality）维度 [56] 、宽度（width）维度 [23]都是可以共存的。因此在固定 尺度（scale）维度之后，我们将调整基数（cardinality）维度和宽度（maintain）维度来保证其复杂度和原始模型类似。我们在本文中没有力求模型大小的最小化，因为这样需要很多如深度可分卷积 [40]、模型剪枝 [19] 和模型压缩 [14] 等细致的方法。

对于 ImageNet [44] 数据集，我们主要使用了 ResNet-50 [23]、ResNeXt-50 [56]、DLA-60 [60] 和 bLResNet-50 [5] 作为基础模型。我们模型的参数量和其原始基础模型大致相同。对于一个 50 层的网络来说，其参数量为 $25M$，一张 224×224 图片的 FLOPs 在 $4.2G$ 左右。对于 CIFAR [27] 数据集，我们主要使用了 ResNeXt-29, 8c×64w [56] 作为基础模型。对于模型复杂度的评估和讨论将在4.4小节给出。

4 实验与分析

4.1 实现细节

我们用 Pytorch 框架实现了我们提出的模型。为了对比的公平，我们也用 Pytorch 实现了 ResNet [23]、ResNeXt [56]、DLA [60] 和 bLResNet-50 [5] 等模型，并且只将其原始的瓶颈块替换为 Res2Net 模块做对比。类似的，我们将在 ImageNet 数据集 [44] 调整过大小的图片中随机剪切出来 224×224 的图片进行训练。我们使用了和 [23], [52] 相同的参数。并且

表 1
ImageNet 数据集上 Top-1 和 Top-5 的错误率

	top-1 错误率 (%)	top-5 错误率 (%)
ResNet-50 [23]	23.85	7.13
Res2Net-50	22.01	6.15
InceptionV3 [52]	22.55	6.44
Res2Net-50-299	21.41	5.88
ResNeXt-50 [56]	22.61	6.50
Res2NeXt-50	21.76	6.09
DLA-60 [60]	23.32	6.60
Res2Net-DLA-60	21.53	5.80
DLA-X-60 [60]	22.19	6.13
Res2NeXt-DLA-60	21.55	5.86
SENet-50 [25]	23.24	6.69
SE-Res2Net-50	21.56	5.94
bLResNet-50 [5]	22.41	-
bLRes2Net-50	21.68	6.00

和 [23] 类似，我们设置优化器为 SGD，$weightdecay = 0.0001$，$momentum = 0.9$，$batchsize = 256$，在 4 块 Titan Xp 上进行训练。初始的学习率设置为 0.1，并且每 30 次迭代将学习率下降为原来的 0.1 倍。

对于 ImageNet 数据集，所有模型（包括基础模型和加入 Res2Net 的模型）都将在相同的参数下迭代 100 次进行训练。对于测试集，我们使用了和 [23] 相同的裁剪策略。对于 CIFAR 数据集，我们实现了 ResNeXt-29 [56]。对所有的任务来说，我们使用了基础模型和只替换瓶颈块为 Res2Net 的新模型。

4.2 ImageNet

我们进行实验所需的数据集 ImageNet [44] 包含了有 1000 种分类标注的 128 万张训练集图片和 5 万张验证集图片。我们构建了大约 50 层的模型来和一流模型进行性能对比。也在 CIFAR 数据集上进行了更多的实验。

4.2.1 模型性能

表1展示出了在 ImageNet 数据集上的 top-1 错误率和 top-5 错误率。简单起见，表1 中的所有 Res2Net 模型的尺度（scale）维度均为 4（$s = 4$）。在 top-1 错误率上，我们的 Res2Net-50 相较于 ResNet-50 有 1.84% 的降低。Res2NeXt-50 的 top-1 错误率也比 ResNeXt-50 降低了 0.85%。并且 Res2Net-DLA-60 的 top-1 错误率比 DLA-60 降低了 1.27%。Res2NeXt-DLA-60 的 top-1 错误率比 DLA-X-60 降低了 0.64%。SE-Res2Net-50 的 top-1 错误率比 SENet-50 降低了 1.68%。bLRes2Net-50 的 top-1 错误率比 bLResNet-50 降低了 0.73%。正如在2.3小节讨论的，即使对于 bLResNet 这种专门设计出来可以利用不同

表 2
更深的网络在 ImageNet 数据集上的 Top-1 和 Top-5 错误率 (%)

	top-1 错误率	top-5 错误率
DenseNet-161 [26]	22.35	6.20
ResNet-101 [23]	22.63	6.44
Res2Net-101	20.81	5.57

尺度特征的网络来说，我们的 Res2Net 模块依然在更细粒度的水平上增强了 bLResNet 的多尺度表达能力。注意，我们使用的 ResNet [23]、ResNeXt [56]、SE-Net [25]、bLResNet [5],、和 DLA [60] 都是现在性能一流的网络。相比较于本身就很优秀的基础框架，集成了 Res2Net 模块的网络依然可以获得性能的提升。

我们也把我们的模型和利用不同大小卷积核并行的 InceptionV3 [52] 进行了比较。公平起见，我们使用了 ResNet-50 [23] 作为基础模型，使用和 InceptionV3 模型一样的 299×299 图象进行训练。我们的 Res2Net-50-299 在 top-1 错误率上比 InceptionV3 降低了 1.14%。因此可以得出结论，我们的层次间类残差连接比 InceptionV3 的并行卷积能够更有效地处理多尺度信息。尽管 InceptionV3 的卷积组合模式是精心设计的，但我们的 Res2Net 模块仍然能简洁而高效的进行模式组合。

4.2.2 更深的 Res2Net

在视觉任务中，更深的网络往往有更好的表达能力 [23], [56]。为了了解我们的模型加深后的表现，我们用均为 101 层的 Res2Net 和 ResNet 进行物体分类性能的比较。如表2所示，我们的 Res2Net-101 在 top-1 错误率上比 ResNet-101 低 1.82%。注意到我们的 Res2Net-50 在 top-1 错误率上是比 ResNet-50 低 1.84% 的。因此，我们的 Res2Net 可以和更深的模型结合，从而拥有更好的表现。同样，我们也对比了 DenseNet [26]。根据官方提供的 DenseNet-161 的 top-1 错误率来看，我们的 Res2Net-101 比它低 1.54%。

4.2.3 尺度（scale）维度的作用

为了验证我们提出尺度（scale）维度的作用，我们实验和分析了有不同的尺度（scale）参数的模型。如表 3所示，更大的尺度（scale）往往有更好的性能。随着尺度（scale）的增加，我们的 Res2Net-50（14w×8s）框架的 top-1 错误率比 ResNet-50 低了 1.99%。为了保证复杂度不变，在尺度（scale）增加的时候，$K_i()$ 的宽度（width）随之减少。我们也进一步证明了，增加尺度（scale）将会增加模型复杂度，但也会提升性能。Res2Net-50（26w×8s）框架比 ResNet-50 的 top-1 错误率低了 3.05%。Res2Net-50（18w×4s）的 top-1 错误率也比 ResNet-50 低了 0.93%，并且 FLOPs 只有其 69%。表3也展示了不同尺度（scales）模型的运行时间，这个表示的是 ImageNet 验证集上 224 × 224 图片的平均耗时。尽管我们需要将特征图分割成 $\{y_i\}$，并且之后需要依次进行层级连接，但是这些额外的运行时间可以被 Res2Net 模型忽略。因为 GPU 能够存储的向量有限，所以当 Res2Net 的 $s = 4$ 时，可以使得 GPU 单周期内高效的并行运算。

表 3
不同尺度（scale）的 Res2Net-50 在 ImageNet 数据集上的 Top-1 错误率和 Top-5 错误率 (%)。
其中 w 是滤波器宽度（width），s 是尺度（scale）数量，参见式1。

	配置	FLOPs	耗时	top-1 错误率	top-5 错误率
ResNet-50	64w	4.2G	149ms	23.85	7.13
Res2Net-50	48w×2s	4.2G	148ms	22.68	6.47
(Preserved	26w×4s	4.2G	153ms	22.01	6.15
complexity)	14w×8s	4.2G	172ms	21.86	6.14
Res2Net-50	26w×4s	4.2G	-	22.01	6.15
(Increased	26w×6s	6.3G	-	21.42	5.87
complexity)	26w×8s	8.3G	-	20.80	5.63
Res2Net-50-L	18w×4s	2.9G	106ms	22.92	6.67

4.3 CIFAR

我们也在 CIFAR-100 [27] 进行了一些实验，这是一个有 100 个分类标注，5 万张图的训练集和 1 万张图的测试集组成的数据集。我们这次使用的基础模型是 ResNeXt-29, 8c×64w [56]。我们需要将原始网络中的基础模块替换为我们的 Res2Net 模块，保持网络其他的配置不变。表4展示了模型的大小和在 CIFAR-100 数据集上的 top-1 错误率。实验结果也表明了，我们的方法比基础模型和其他方法的参数要少，并且性能更优。我们的 Res2NeXt-29, 6c×24w×6s 比基础模型的 top-1 错误率低 1.11%。Res2NeXt-29, 6c×24w×4s 的参数量甚至只有 ResNeXt-29, 16c×64w 的 35%。对比 DenseNet-BC ($k = 40$)，我们也实现了一个用更少参数来获得更高性能的模型。相比较于 Res2NeXt-29, 6c×24w×4s，Res2NeXt-29, 8c×25w×4s 因为有更大的宽度（width）和基数（cardinality），所以获得了更好的性能。这也表明了尺度（scale）维度和宽度（width）、基数（cardinality）是可以共存的。我们也将最近新提出的 SE 模块集成在了我们的结构中。我们的方法可以使用比基础模型 ResNeXt-29, 8c×64w-SE 更少的参数获得更高的性能。

4.4 改变尺度（scale）维度

类似于 Xie 等人 [56]，我们改变网络的不同维度的数值来测试其性能，这些维度包括尺度（scale）(如式1)、基数（cardinality）[56] 和深度（depth）[47]。当我们增加一个模型的一个维度的时候，会保持其他维度不变。这一系列网络将在上述条件的改变下被训练和检测。因为 [56] 中已经验证了增加基数

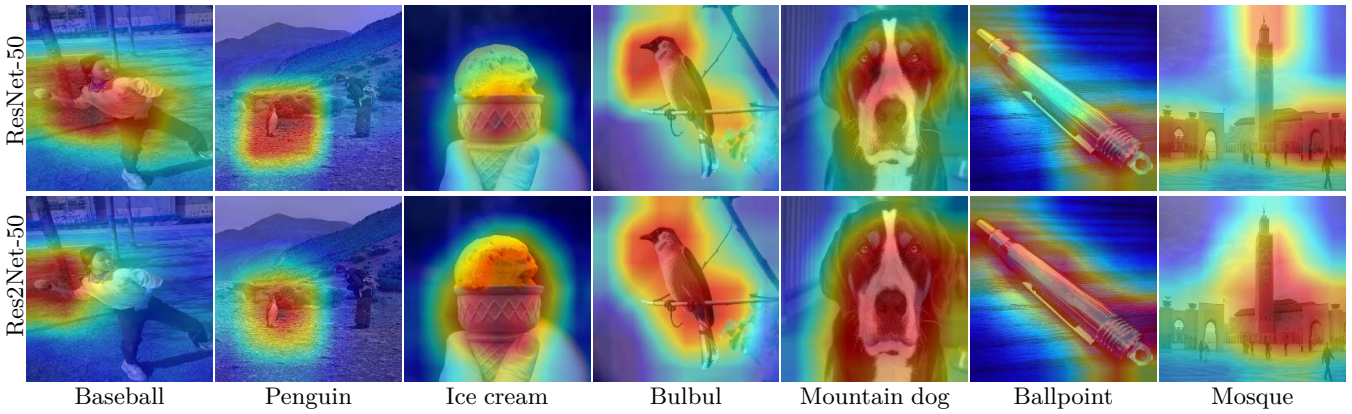


图 4. ResNet-50 和 Res2Net-50 的 CAM [45] 可视化效果

表 4
在 CIFAR-100 数据集上的 Top-1 错误率 (%) 和模型大小
参数 c 表示基数的值，w 表示滤波器宽度

	参数量	top-1 错误率
Wide ResNet [61]	36.5M	20.50
ResNeXt-29, 8c×64w [56] (base)	34.4M	17.90
ResNeXt-29, 16c×64w [56]	68.1M	17.31
DenseNet-BC (k = 40) [26]	25.6M	17.18
Res2NeXt-29, 6c×24w×4s	24.3M	16.98
Res2NeXt-29, 8c×25w×4s	33.8M	16.93
Res2NeXt-29, 6c×24w×6s	36.7M	16.79
ResNeXt-29, 8c×64w-SE [25]	35.1M	16.77
Res2NeXt-29, 6c×24w×4s-SE	26.0M	16.68
Res2NeXt-29, 8c×25w×4s-SE	34.0M	16.64
Res2NeXt-29, 6c×24w×6s-SE	36.9M	16.56

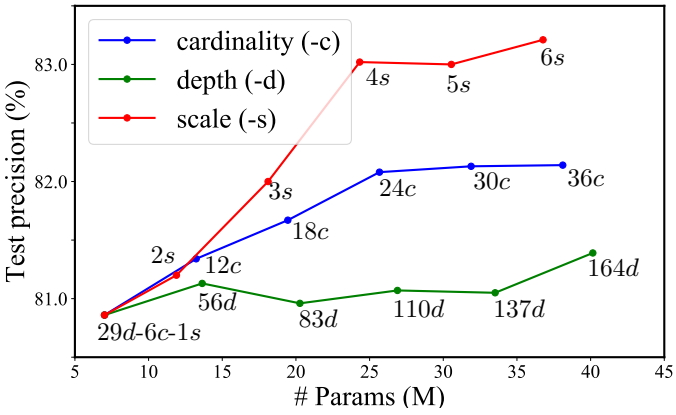


图 5. CIFAR-100 数据集上的不同大小和参数的模型测试，
改变了基数（ResNeXt-29）、深度（ResNeXt）和尺度（Res2Net-29）。

表 5
在 PASCAL VOC07 和 COCO 数据集上的物体检测结果，
使用 AP (%) 和 AP@IoU=0.5 (%) 作为测试标准。
Res2Net 和其对比的网络有着相似的复杂度。

数据集	骨架结构	AP	AP@IoU=0.5
VOC07	ResNet-50	72.1	-
	Res2Net-50	74.4	-
COCO	ResNet-50	31.1	51.4
	Res2Net-50	33.7	53.6

（cardinality）比增加宽度（width）更加有效，所以我们只对比了尺度（scale）维度与基数（cardinality）、深度（depth）。

图5展示了 CIFAR-100 数据集上的不同大小和参数的模型测试结果。基础模型的深度、基数和尺度分别是 29、6 和 1。实验结果表明了尺度（scale）对于模型的性能很重要，这也和4.2 小节中在 ImageNet 数据集上的实验结果相吻合。并且增加尺度（scale）也比增加其他维度能更快的提升网络性能。如式 1和图2所示，在尺度 $s = 2$ 的情况下，我们只是增加了模型中 1×1 滤波器的参数而已。因此，模型中 $s = 2$ 时，其性能会比增加基数（cardinality）略差。对于 $s = 3, 4$ 来说，我们的层次间类残差连接结构能够产生一系列丰富的等效尺度集合，这有利于获得更好的性能。不过当尺度是 5 和 6 时，就只能获得有限的性能提升，这可能是因为 CIFAR 数据集的图象太小（32×32），没有丰富的多尺度信息。

4.5 CAM 图

为了更好地理解 Res2Net 的多尺度表达能力，我们使用 Grad-CAM [45] 方法可视化了 CAM 图，这种方法是用来定位图像分类器的敏感区域的。如图4的可视化示例中，越强的 CAM 区域使用了越亮的颜色。我们对比了 Res2Net 和 ResNet 在棒球，企鹅等小物体上的表现。这两种框架对类似于冰淇淋等中等大小的物体都有着相似的 CAM 图表现。由于有着更好的多尺度表达能力能力，Res2Net 的 CAM 图更倾向于覆盖整个物体，而 ResNet 则只覆盖物体的一部分，如图中的夜莺、山狗、圆珠笔和清真寺。这种能够在 CAM 图中精准定位物体所在区域的能力，对于弱监督语义分割有着可挖掘的潜在价值 [54]。

表 6
在 COCO 数据集上，其在不同尺寸物体上的 AP 和 AR 表现。

		物体尺寸			
		Small	Medium	Large	All
ResNet-50	AP (%)	13.5	35.4	46.2	31.1
Res2Net-50		14.0	38.3	51.1	33.7
Improve.		+0.5	+2.9	+4.9	+2.6
ResNet-50	AR (%)	21.8	48.6	61.6	42.8
Res2Net-50		23.2	51.1	65.3	45.0
Improve.		+1.4	+2.5	+3.7	+2.2

表 7
在 PASCAL VOC12 数据集上，使用不同尺度（scale）的 Res2Net-50 的表现。
Res2Net 和其对比模型有着相似的复杂度。

骨架结构	配置	Mean IoU (%)
ResNet-50	64w	77.7
Res2Net-50	48w×2s	78.2
	26w×4s	79.2
	18w×6s	79.1
	14w×8s	79.0
ResNet-101	64w	79.0
Res2Net-101	26w×4s	80.2

4.6 物体检测

对于物体检测这个任务，我们使用了 Faster R-CNN [43] 作为基础模型，在 PASCAL VOC07 [16] 数据集和 MS COCO [33] 数据集上验证了我们的 Res2Net 。我们使用了 ResNet-50 和 Res2Net-50 作为骨架网络进行对比，并且公平起见，其他实现细节也都设计的一样。表5中展示了物体检测的结果。在 PASCAL VOC07 数据集上，Res2Net-50 模型比其对比的模型 AP 变优了 2.3%。在 COCO 数据集上，Res2Net-50 模型比其对比的模型 AP 变优了 2.6%，AP@IoU=0.5 变优了 2.2%。

我们也测试模型在不同尺寸物体上的 AP 和 AR，表6为测试结果。根据 [33] 的标准，物体按照尺寸不同被分为三种。Res2Net 模型比其他基础模型取得了很大的进步。在小尺寸，中尺寸，大尺寸的物体的 AP 分别提升了 0.5%、2.9% 和 4.9%，AR 分别提升了 1.4%、2.5% 和 3.7% 。因为有着更强的多尺度表达能力，Res2Net 模型可以用更大的感受野来覆盖物体，这样也提升了其在不同尺寸物体上的表现。

4.7 语义分割

语义分割需要卷积神经网络对物体的上下文语境信息有很强的多尺度提取能力。因此我们验证了 Res2Net 在 PASCAL VOC12 数据集上进行语义分割的表现。我们使用的 PASCAL VOC12 数据集 [20] 由包含 10582 张图片的训练集和包含 1449 张图片的测试集组成。本次使用的基础模型是 Deeplab v3+ [8]。除了将骨干网络替换为 ResNet 和 Res2Net 之外，其他保持和 Deeplab v3+ [8] 相同。在训练和验证时使用的步长（strides）都是 16。如表7所示，Res2Net-50 比他的对比模型在 mean IoU 上提升了 1.5%。Res2Net-101 比他的对比模型在 mean IoU 上提升了 1.2%。图6中也将部分语义分割结果进行了可视化。Res2Net 模型倾向于对任何尺寸物体的所有部分都进行覆盖。

表 8
不同尺度（scale）的 Res2Net-50 在 COCO 数据集上实例分割的性能表现。Res2Net 将和其对比模型复杂度近似。

骨节结构	配置	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50	64w	33.9	55.2	36.0	14.8	36.0	50.9
Res2Net-50	48w×2s	34.2	55.6	36.3	14.9	36.8	50.9
	26w×4s	35.6	57.6	37.6	15.7	37.9	53.7
	18w×6s	35.7	57.5	38.1	15.4	38.1	53.7
	14w×8s	35.3	57.0	37.5	15.6	37.5	53.4

4.8 实例分割

实例分割是物体检测和语义分割的结合。他不仅需要识别出各种尺寸的物体，也要准确的分割出每个物体。正如在4.6小节和4.7小节中分析的，物体检测和语义分割都需要卷积神经网络有很强的多尺度表达能力。因此，实例分割将能从更优的多尺度表达能力上获益。我们使用了 Mask R-CNN [21] 作为实例分割的算法，我们只将其骨架由 ResNet-50 替换为 Res2Net-50。在 MS COCO [33] 数据集上的实例分割表现如表8所示。Res2Net-26w×4s 模型比其对比模型 AP 变优了 1.7%，AP_{50} 变优了 2.4%。其也展示了对于不同的尺寸物体的性能提升。对于小尺寸、中尺寸、大尺寸物体，其 AP 提升分别是 0.9%、1.9% 和 2.8%。表8中也展示了 Res2Net 在相同复杂度不同尺度（scale）下的性能对比。随着尺度（scale）的增加，性能朝着上升的趋势发展。注意到 Res2Net-50-26w×4s 相较于 Res2Net-50-48w×2s 在 AP_L 性能上提升了 2.8 %，同时 Res2Net-50-48w×2s 和 ResNet-50 有着相同的 AP_L。我们猜想对于大的物体，模型的性能因为额外的尺度（scale）提升。当尺度（scale）相对较大时，性能提升将不明显。Res2Net 模型能够自适应性的调整感受野范围。当整个图象中的物体的已经被合适的感受野覆盖时，模型的性能提升将变得有限。当模型的复杂度不变的情况下，单纯增加模型尺度（scale）可能造成每个感受野的通道数减少，这可能会降低模型对于特定尺度特征的处理能力。

4.9 显著性物体检测

像是显著性物体检测这种像素水平的视觉任务，也需要卷积神经网络有很强的多尺度表达能力来定位整个物体和其区域

图 6. 使用 ResNet-101 和 Res2Net-101 作为模型骨架，其语义分割结果的可视化 [8]。

表 9
在不同数据集上的显著性物体检测结果，评判标准使用 F-measure 和 MAE。保证 Res2Net 和其对比模型的复杂度类似。

数据集	骨干结构	F-measure↑	MAE ↓
ECSSD	ResNet-50	0.910	0.065
	Res2Net-50	0.926	0.056
PASCAL-S	ResNet-50	0.823	0.105
	Res2Net-50	0.841	0.099
HKU-IS	ResNet-50	0.894	0.058
	Res2Net-50	0.905	0.050
DUT-OMRON	ResNet-50	0.748	0.092
	Res2Net-50	0.800	0.071

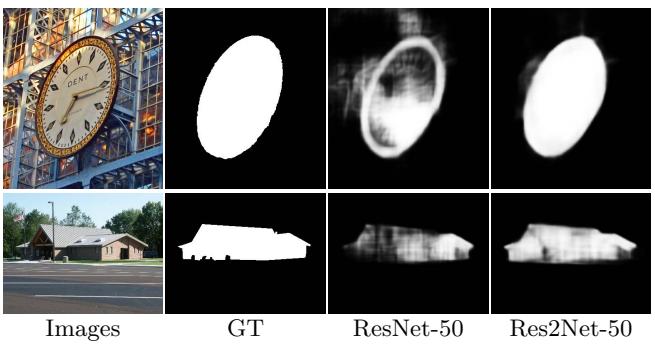


图 7. 使用 ResNet-50 和 Res2Net-50 作为骨干结构的一些显著性物体检测 [24] 结果示例。

表 10
COCO 验证数据集上关键点预测的性能表现。Res2Net 和其对比模型的复杂度相似。

骨干结构	AP	AP_{50}	AP_{75}	AP_M	AP_L
ResNet-50	70.4	88.6	78.3	67.1	77.2
Res2Net-50	73.7	92.5	81.4	70.8	78.2
ResNet-101	71.4	89.3	79.3	68.1	78.1
Res2Net-101	74.4	92.6	82.6	72.0	78.5

边界。我们使用了最新的 DSS [24] 作为我们的基础模型。公平起见，也只将其骨干结构替换为 ResNet-50 和 Res2Net-50，同时其他配置参数保持不变。如 [24]，我们使用 MSRA-B 数据集 [35] 进行训练，在 ECSSD [58]、PASCAL-S [30]、HKU-IS [29] 和 DUT-OMRON [59] 数据集上验证结果。我们使用 F-measure 和 MAE 作为检测标准，如表 9所示，集成了 Res2Net 的模型相较于其他模型性能均有提升。在 DUT-OMRON 数据集（包含 5168 张图片）上，集成 Res2Net 的模型比集成 ResNet 的在 F-measure 上优 5.2%，在 MAE 上优 2.1%。我们的 Res2Net 方法在 DUT-OMRON 数据集上的性能提升最大，因为这个数据集相较于其他数据集，其图像中的重要物体大小变化会更大。如图7所示，是一些显著性物体检测样例的可视化对比。

4.10 关键点预测

人体的每个部分尺寸都不相同，因此我们需要一个好的关键点预测算法来定位人体不同大小的身体部位。为了验证 Res2Net 的多尺度表达能力是否对关键点预测这个任务有效用，我们使用了 SimpleBaseline [55] 作为基础模型，只将其骨干结构替换为 Res2Net。包括训练和测试在内的所有实现方法都和 SimpleBaseline [55] 保持一致。我们的训练集使用 COCO 关键点检测数据集 [33]，测试集使用 COCO 验证数据集。跟 SimpleBaseline [55] 中的一样，我们使用了相同的人体检测器。表 10中展示了我们的 Res2Net 使用 COCO 验证集进行测试时的性能表现。测试结果表明，Res2Net-50 和 Res2Net-101 模型的 AP 为 3.3% 和 3.0%。并且，Res2Net 模型相较于其基准模型，在不同尺度的人体上都有一个不错的性能提升。

5 总结

我们提出了一种可将卷积神经网络的多尺度表达能力提升到更细粒度层次的简洁而高效的模型，我们将之命名为Res2Net。Res2Net 扩展出了一个名叫尺度（scale）的维度，这个维度比现存的深度（depth），宽度（width），基数（cardinality）等维度要更加重要且有效。我们的 Res2Net 模块也可以毫不费力的集成在现有的一流模型上。在 CIFAR-100 和 ImageNet 两个数据集的图像分类任务中，我们的模型也比包括 ResNet、ResNeXt、DLA 等模型在内的其他一流模型有更好的性能。

尽管在包括 CAM、物体检测、显著性物体识别这几个有代表性的视觉任务在内的很多任务中，我们已经证明了我们提出的框架的优越性，但我们认为这种强大的多尺度表达能力将有更广泛的应用范围。为了鼓励后来学者利用这种强大的多尺度表达能力进行研究，我们将代码在https://mmcheng.net/res2net/进行开源。

致谢

研究受到 NSFC (NO. 61620106008, 61572264)，国家”万人计划”青年拔尖人才支持计划，天津市自然科学基金 (17JCJQJC43700, 18ZXZNGX00110) 的支持。

参考文献

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell., 24(4):509–522, 2002.

[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. Computational Visual Media, 5(2):117–150, 2019.

[3] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. IEEE Trans. Image Process., 24(12):5706–5722, 2015.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In IEEE Conf. Comput. Vis. Pattern Recog., pages 1021–1030, 2017.

[5] C.-F. R. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris. Big-Little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition. In Int. Conf. Mach. Learn., 2019.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell., 40(4):834–848, 2018.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. CoRR, abs/1706.05587, 2017.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In The European Conference on Computer Vision (ECCV), September 2018.

[9] Y. Chen, H. Fang, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In Int. Conf. Comput. Vis., 2019.

[10] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. In Adv. Neural Inform. Process. Syst., pages 4467–4475, 2017.

[11] B. Cheng, R. Xiao, J. Wang, T. Huang, and L. Zhang. High frequency residual learning for multi-scale image classification. In Brit. Mach. Vis. Conf., 2019.

[12] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. S. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. Computational Visual Media, 5(1):3–20, Mar 2019.

[13] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell., 37(3):569–582, 2015.

[14] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. A survey of model compression and acceleration for deep neural networks. CoRR, abs/1710.09282, 2017.

[15] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In IEEE Conf. Comput. Vis. Pattern Recog., July 2017.

[16] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis., 88(2):303–338, 2010.

[17] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell., pages 1–1, 2020.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 580–587, 2014.

[19] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In Adv. Neural Inform. Process. Syst., pages 1135–1143, 2015.

[20] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In Int. Conf. Comput. Vis. IEEE, 2011.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In Int. Conf. Comput. Vis., pages 2961–2969, 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell., 37(9):1904–1916, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., pages 770–778, 2016.

[24] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell., 41(4):815–828, 2019.

[25] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[27] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Adv. Neural Inform. Process. Syst., pages 1097–1105, 2012.

[29] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5455–5463, 2015.

[30] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 280–287, 2014.

[31] M. Lin, Q. Chen, and S. Yan. Network in network. In Int. Conf. Learn. Represent., 2013.

[32] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In IEEE Conf. Comput. Vis. Pattern Recog., volume 1, page 4, 2017.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In Eur. Conf. Comput. Vis., pages 740–755. Springer, 2014.

[34] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang. A simple pooling-based design for real-time salient object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3917–3926, 2019.

[35] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell., 33(2):353–367, 2011.

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In Eur. Conf. Comput. Vis., pages 21–37. Springer, 2016.

[37] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang. Richer convolutional features for edge detection. IEEE Trans. Pattern Anal. Mach. Intell., 41(8):1939 – 1946, 2019.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3431–3440, 2015.

[39] D. G. Lowe. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis., 60(2):91–110, 2004.

[40] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Eur. Conf. Comput. Vis., September 2018.

[41] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In Proceedings of the IEEE International Conference on Computer Vision, pages 4875–4884, 2017.

[42] G.-Y. Nie, M.-M. Cheng, Y. Liu, Z. Liang, D.-P. Fan, Y. Liu, and Y. Wang. Multi-level context ultra-aggregation for stereo matching. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3283–3291, 2019.

[43] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Adv. Neural Inform. Process. Syst., pages 91–99, 2015.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis., 115(3):211–252, 2015.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Int. Conf. Comput. Vis., pages 618–626, 2017.

[46] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Adv. Neural Inform. Process. Syst., pages 568–576, 2014.

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Int. Conf. Learn. Represent., 2014.

[48] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5693–5703, 2019.

[49] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. High-resolution representations for labeling pixels and regions. CoRR, abs/1904.04514, 2019.

[50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. volume 4, page 12, 2017.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1–9, 2015.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2818–2826, 2016.

[53] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng. Salient object detection: A discriminative regional feature integration approach. Int. J. Comput. Vis., 123(2):251–268, 2017.

[54] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[55] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In Eur. Conf. Comput. Vis., September 2018.

[56] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5987–5995. IEEE, 2017.

[57] S. Xie and Z. Tu. Holistically-nested edge detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1395–1403, 2015.

[58] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1155–1162, 2013.

[59] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3166–3173, 2013.

[60] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. Deep layer aggregation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2403–2412, 2018.

[61] S. Zagoruyko and N. Komodakis. Wide residual networks. In Brit. Mach. Vis. Conf., pages 87.1–87.12, September 2016.

[62] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 202–211, 2017.

[63] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[64] J. Zhao, Y. Cao, D.-P. Fan, X.-Y. Li, L. Zhang, and M.-M. Cheng. Contrast prior and fluid pyramid integration for rgbd salient object detection. In IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[65] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng. Optimizing the F-measure for threshold-free salient object detection. In Int. Conf. Comput. Vis., 2019.

[66] K. Zhao, W. Shen, S. Gao, D. Li, and M.-M. Cheng. Hi-Fi: Hierarchical feature integration for skeleton detection. In Int.

Joint Conf. Artif. Intell., 2018.
[67] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1265–1274, 2015.

Shang-Hua Gao is a master student in Media Computing Lab at Nankai University. He is supervised via Prof. Ming-Ming Cheng. His research interests include computer vision, machine learning, and radio vortex wireless communications.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012, and then worked with Prof. Philip Torr in Oxford for 2 years. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer vision and computer graphics. He received awards including ACM China Rising Star Award, IBM Global SUR Award, etc. He is a senior member of the IEEE and on the editorial boards of IEEE TIP.

Kai Zhao Kai Zhao is currently a Ph.D candidate with college of computer science, Nankai University, under the supervision of Prof Ming-Ming Cheng. His research interests mainly focus on statistical learning and computer vision.

Xin-Yu Zhang is an undergraduate student from School of Mathematical Sciences at Nankai University. His research interests include computer vision and deep learning.

Ming-Hsuan Yang is a professor in Electrical Engineering and Computer Science at University of California, Merced. He received the PhD degree in Computer Science from the University of Illinois at Urbana-Champaign in 2000. Yang has served as an associate editor of the IEEE TPAMI, IJCV, CVIU, etc. He received the NSF CAREER award in 2012 and the Google Faculty Award in 2009.

Philip Torr received the PhD degree from Oxford University. After working for another three years at Oxford, he worked for six years for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He is now a professor at Oxford University. He has won awards from top vision conferences, including ICCV, CVPR, ECCV, NIPS and BMVC. He is a senior member of the IEEE and a Royal Society Wolfson Research Merit Award holder.