ISCAS 中国科学院软件研究所学术年会'2021
暨计算机科学国家重点实验室开放周

学术论文

用于高效视觉任务的多用卷积核

（Minor revision in T-PAMI）

韩凯 吴恩华

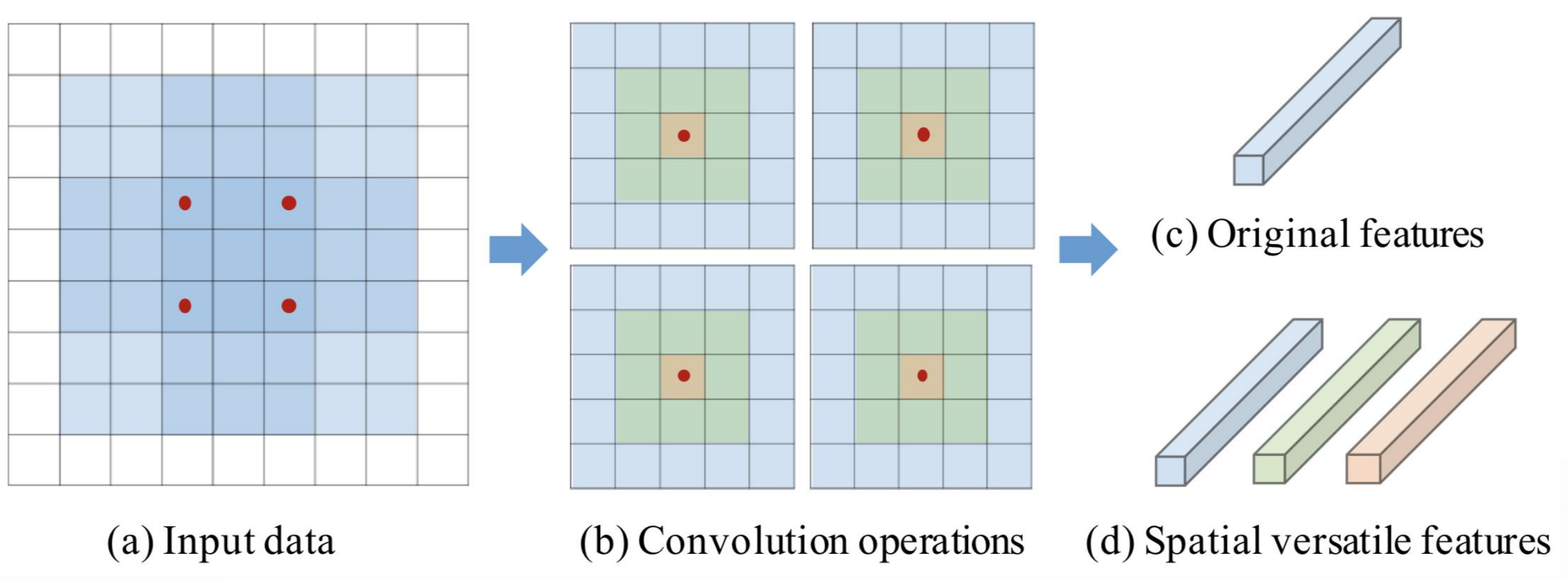
摘要

卷积神经网络模型在计算机视觉、图形学等领域应用广泛，但是卷积网络体积大速度慢。本文提出用于加速神经网络的多用卷积核，利用少量卷积核生成更多卷积核。所提出多用卷积核能够大幅降低卷积网络的计算量和参数量。

方法

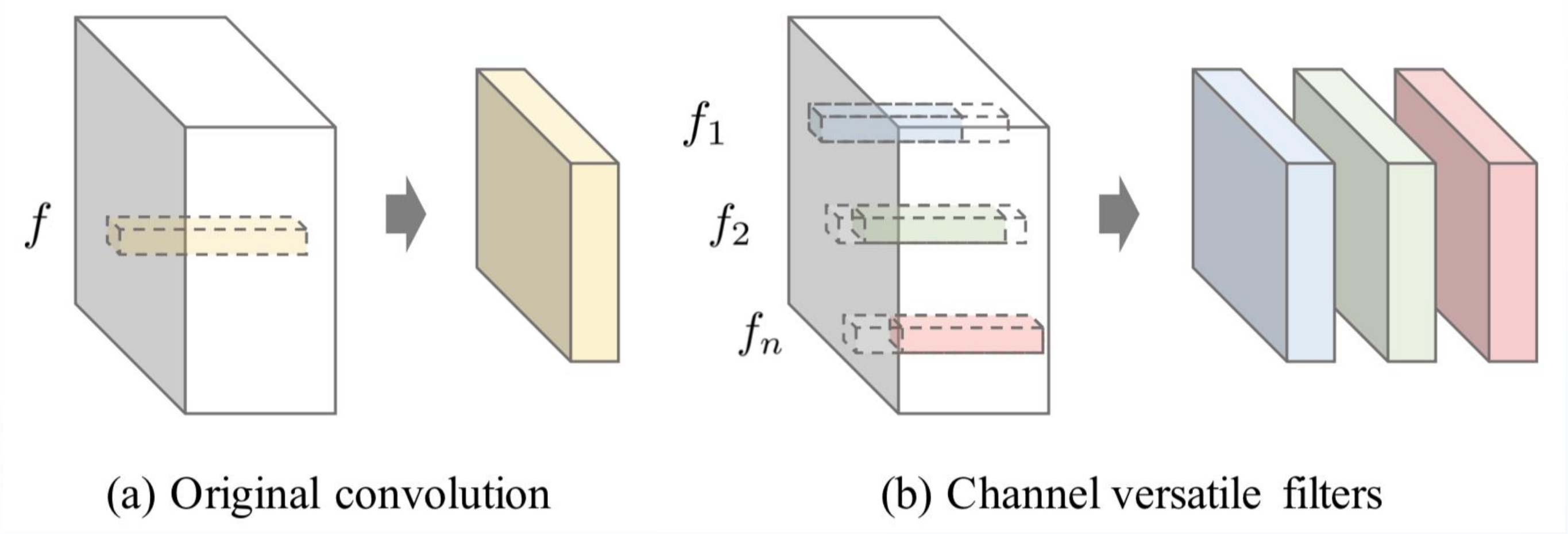
建立少量主卷积核，从主卷积核中产生更多子卷积核。

1. 空间多用卷积核：从高度、宽度维度“挖”出子卷积核。
2. 通道多用卷积核：从通道维度“挖”处子卷积核。
3. 掩码可学多用卷积核：不限制维度，让二值掩码是可学习的，和权重一起训练。



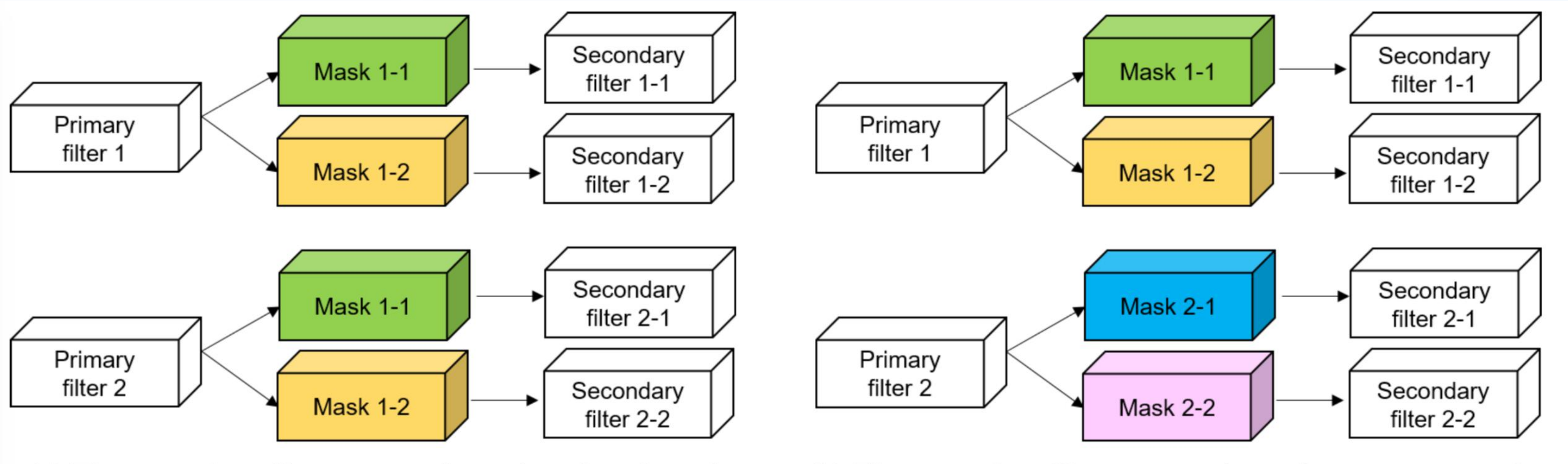
(a) Input data (b) Convolution operations (d) Spatial versatile features

空间多用卷积核



(a) Original convolution (b) Channel versatile filters

通道多用卷积核



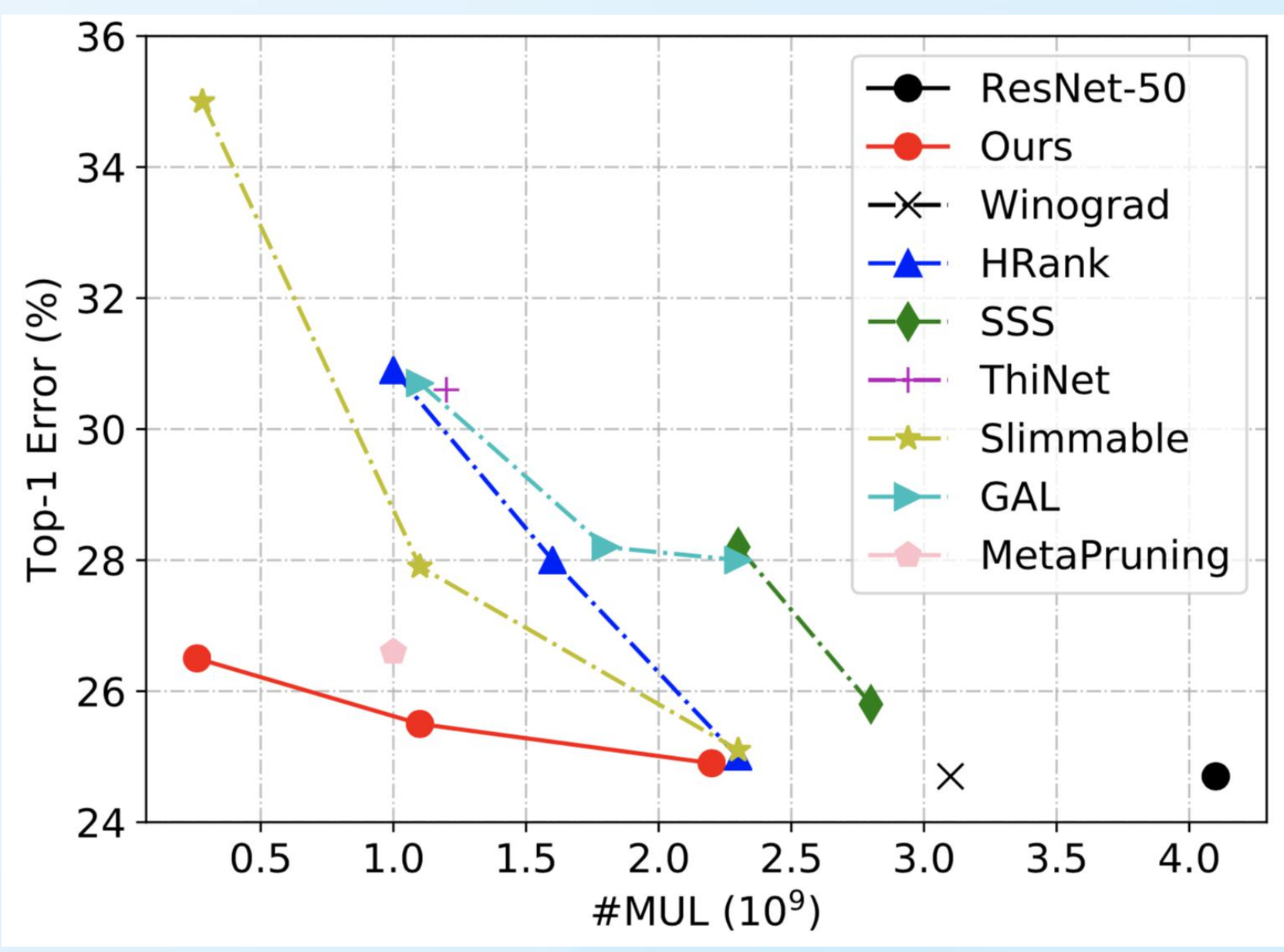
(a) The secondary filters generation using shared masks. (b) The secondary filters generation using separate masks.

掩码可学多用卷积核

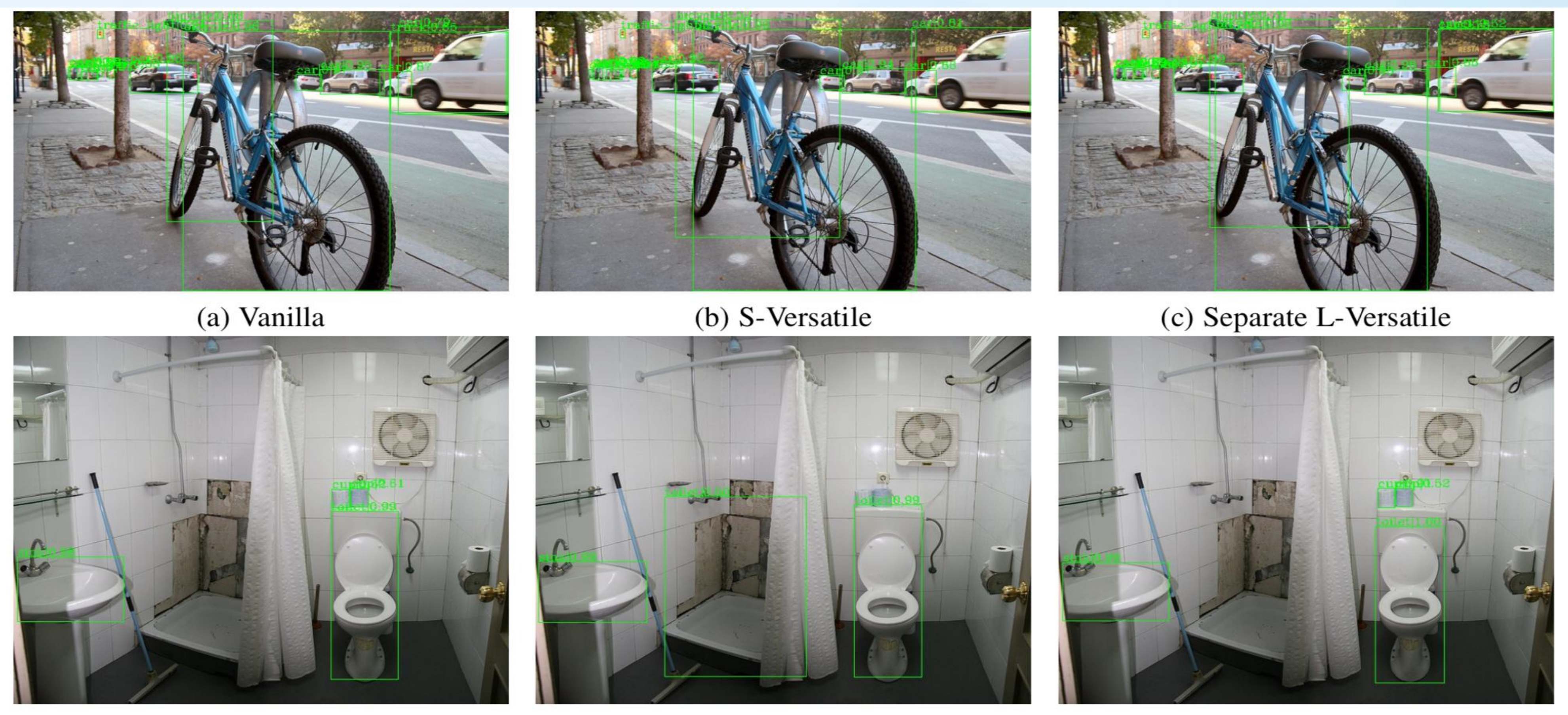
实验

ImageNet图像分类模型压缩

ResNet-50	#Param $(\times 10^7)$	Mem (MB)	#MUL $(\times 10^9)$	Top1err	Top5err
Vanilla [20]	2.6	97	4.1	24.7%	7.8%
S-Versatile	1.9	76	3.0	24.5%	7.6%
S+C-Versatile	1.1	42	1.5	25.5%	8.2%
Shared L-Versatile (s=4)	0.8	30	1.1	26.8%	8.9%
Separate L-Versatile (s=4)	0.9	33	1.1	25.5%	8.0%
Separate L-Versatile (s=32)	0.36	14	0.26	26.5%	8.7%



COCO目标检测模型压缩



(a) Vanilla (b) S-Versatile (c) Separate L-Versatile

(d) Vanilla (e) S-Versatile (f) Separate L-Versatile

联系方式：韩凯 hankai@ios.ac.cn 13621227542