通过分层蒸馏实现的高效实例分割

摘要

最近实例分割在公共基准上取得了大的精准度的提升。然而，由于这些模型推理速度较慢，难以用于实际场景。本文中，我们提出了一种利用师生学习框架的更快的实例分割模型，该模型通过将训练好的老师模型中获得的知识转移到轻量的学生模型中。除了都是单任务网络的分类或语义分割中传统的知识提取策略之外，我们还研究了基于多任务学习的实例分割中结构信息蒸馏的分层蒸馏框架（H-Dis）。H-Dis 由两种蒸馏方案构成：特征蒸馏，蒸馏由多头共享的成对的量化特征图，语义蒸馏，确保蒸馏实例级别的每个头部信息。特别是，我们为分割头部提出通道级别的蒸馏，实现实例水平的掩模（mask）知识转移。为了评估我们的方法，我们在不同的数据集 Pascal VOC 和 Cityscapes 中实验不同的蒸馏设置方法。我们的实验证明，在有限的计算资源下，我们的方法可以有效的加速实例分割模型，且精度损失较小。

引言

近年来，深度卷积网路极大地提升了实例分割的性能，这为将其应用于实际应用提供了诱人的可能性，如监视系统，自动驾驶系统和医疗系统。通常，更强大的 DNN 模型具有更深更大的网络设计，这使其推理速度较低，与此相反，速度是绝大多数实际应用的关键需求。因此，迫切需要一个更快更轻的高精度的实例分割模型。

先前的研究人员发现了一些加速 DNN 模型的方法。模型压缩【7，14，27，28】分解每一层的权重以消除冗余，通过逐层重构和微调来恢复一定的准确性。模型剪枝选择具有较高重要性或稀疏约束的层通道。这些方法可以显著提升速度，但是准确性下降也很明显，尤其对复杂任务的目标检测和分割。

为了提高微小模型或压缩模型的准确度，知识转移是一种很好的方式，将笨重的老师模型学习到的有用有效的知识转移到轻量的学生模型。将知识从一个模型转移到另一个模型流行的方法是知识蒸馏和模仿学习。常规地将知识蒸馏用于分类网络。训练紧凑的语义分割，其问题可以简单的看作逐像素分类，知识蒸馏也可以直接用于逐像素蒸馏。然而，分类蒸馏和语义分割蒸馏都是单任务蒸馏。在流行的多任务实例分割网络中应用知识蒸馏是一项挑战，因为是实例级别的多类语义信息的结构信息传递，分类和语义分割都不在这个范围。

本文中，我们为实例分割提出了分层蒸馏框架，中间层输出作为特征蒸馏，后期层输出作为语义蒸馏将输出蒸馏，将结构信息从笨重的教师网络转移到紧凑的学生信息。H-Dis 是基于 RoI 操作，可优化实例级别蒸馏。特征蒸馏在 RoI 池化后每个头部网络的量化特征图来蒸馏。语义蒸馏使用每个头部网络的输出，包括分类，矩形框回归，掩模预测。与在语义分割中使用 softmax 的像素水平分类不同，实例分割的掩模分支设计为不存在类间竞争（使用 sigmoid 替代 softmax），这比 softmax 获得更大的收益。因此，我们为掩模分支提出逐通道蒸馏，此外还有蒸馏分类网络得到的软目标和边界框回归网络的边界框坐标的直接方案。

总之，本文贡献有三点：

我们为实例分割提出了新的分层蒸馏框架，利用教师网络的中间层和后期层输出作为我们的蒸馏目标，以快速分割的速度训练一个高效的学生网络。据我们所知，这是对实例分割首次尝试知识蒸馏。

我们为多任务实例分割框架的掩模分支设计了新的蒸馏损失，为师生学习使用逐通道蒸馏。

在大型基准测试中评估证明了提出方法的有效性。

相关工作

实例分割：任务是为图像中的每个目标分配实例级别的分割掩码。通常，该任务需要分割，分类和框回归，这些可以分别或联合基于区域的方法，例如候选区域网络（RPN）来完成。这种范例包括最流行的方法，如 SDS，CFM 和 MNC。最近，FCIS 被提出作为一种完整的端

到端全卷积网络用于实例分割，同时联合执行位置敏感通道的掩膜估计，分类和回归输出，达到了高精度和快速度。Mask R-CNN 通过在 Faster R-CNN 中添加 FCN 分支扩展，该分支用于和类与框预测分支并行预测目标掩码。Mask R-CNN 在公共基准上取得了竞争性的精度，但是速度不在他们的考虑之内。考虑速度，精度和训练时间，我们在此工作中使用原始的 FCIS 作为基准。

知识转移：知识蒸馏的首创工作，Hinton 等人提出了一种有用的方法显著提升了小模型的精度，通过转移网络整体的泛化能力，显著增强了图像分类工作的性能。这个想法是让学生网络不仅捕获真实标签提供的信息，还包括教师网络学习到的更精细结构的额外信息。随后的工作试图通过转移中间特征来解决其中的缺点。Romero 等进一步开发了这个想法，使用更窄更深的学生网络模仿一个宽而浅的教师网络的全部特征图。然而，由于教师和学生的能力相差太大，这个假设太严格了。在某些情况下，对性能和收敛有不利的影响。Chen 等提出了用于分类和框回归的蒸馏损失，并在目标检测网络中应用了模拟学习。这项工作提出了将知识迁移用于多任务网络的一般方法。然而，这项工作只能处理目标检测问题，在优化基于 RoI 的任务的模拟学习时没有贡献。Li 等扩展了目标检测任务的模拟学习，仅在 RoIs 中传递特征来解决问题。Xie 等研究了语义分割网络的知识蒸馏，在掩码上应用像素级的蒸馏，并通过一致性蒸馏来得到掩码边缘信息。Liu 等进一步提出了用于 GAN 的整体蒸馏技术，匹配老师和学生产生的掩码。这些工作为分割蒸馏提供了可行的方法，但是仅限于单任务语义分割网络。

方法

我们从繁琐的教师网络蒸馏知识到轻且高效的学生网络。它们都遵循最高水平的实例架构，在网络头部有分类，框回归，掩膜分支。教师网络分割精度的性能优于学生网络，但是运行速度更慢。正如【2】所做的那样，建议在头部网络和共享卷积层都添加知识蒸馏。

我们的方法采用教师网络的中层和后层输出作为我们的蒸馏目标，将结构信息转移到学生网络。我们提出的方法概述如图 1。

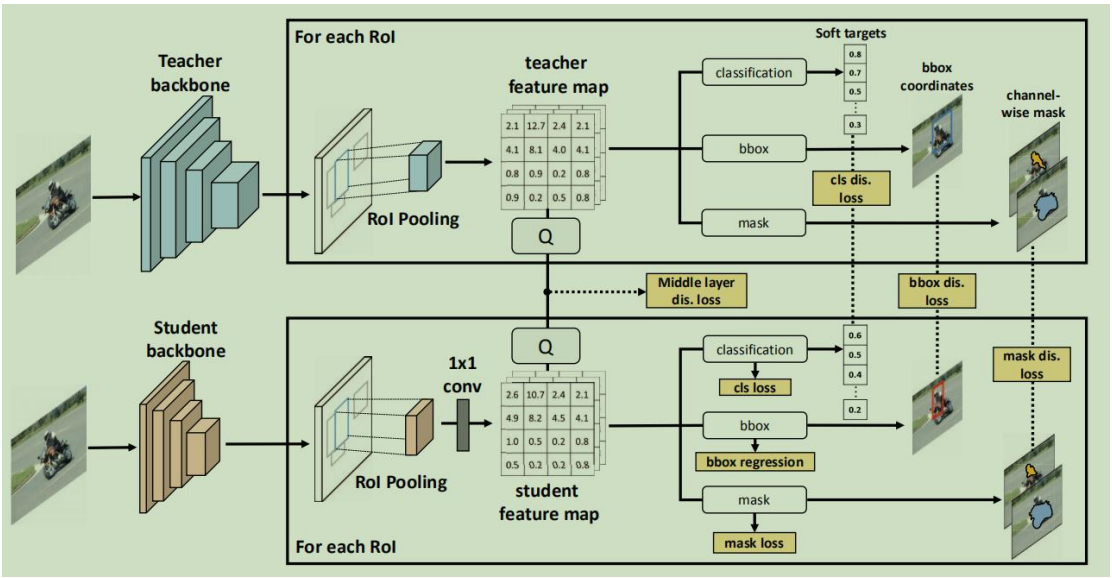

图 1：我们提出的实例分割分层蒸馏架构概述。首先，我们使用所有的训练数据训练教师网络，在蒸馏阶段只前向传播输入图片。其次，我们在教师网络和学生网络 RoIs 上的特征图用 L2 回归使用特征蒸馏作为中间层的蒸馏损失。由于大多数多任务网络更重更深，因此在特征图上进行量化操作使学生网络从老师网络学习收敛更快更简单。最后，我们还从教师网络蒸馏了后期层输出作为语义信息，但为每个头部定制了，包括每个 RoI 的软目标分类，边界框坐标，每个通道掩码。语义蒸馏损失由类别蒸馏的交叉熵损失和方框与掩码蒸馏的 L2

损失组成。

特征蒸馏

对实例分割，从共享的卷积层提取特征图会影响定位，分类和分割的精度。将包含许多的黑箱知识的中间层转移来促进学生网络是必要的。实例分割是基于区域的工作，头部网络基于池化的候选区域，所以在此任务中区域候选有非常重要的作用。从中间层蒸馏知识，我们遵循【23】，在教师网络和学生网络上对特征图引入量化操作。量化方法是教师网络和学生网络的输出离散化，使学生网络能更好的匹配教师网络。特征蒸馏损失函数 L_{rp_dis} 定义如下：

$$L_{rp\_dis} = \frac{1}{2N}\sum_{i\in N}\|Q(f_t^i) - Q(r(f_s^i))\|_2^2$$

Q 是继承自【23】，元素级别的量化函数。N 是提案数目，r 是将学生特征图转为教师特征图同样大小的函数。这里，我们使用 1x1 的卷积层作为 r。

语义蒸馏

我们采用网络中头部的输出，这些输出是各种各样语义信息。在流行的实例分割框架中有三个分支，包括分类，框回归，掩码预测。由于头部网络是基于 RPN 生成的候选区域训练的，因此用于训练学生网络的候选区域和教师网络一样。

分类和框回归蒸馏。训练分类网络，已经提出了常规的知识蒸馏，将老师模型预测的类别作为“软目标”来指导学生网络的训练。神经网络通常使用 softmax 输出层来产生类别概率，其中 softmax 输出层将预测第 i 个类别得分的输出 zi，计算转为其类别概率 pi，将 zi 与其它的对数变换 pi=softmax（zi/T）比较。T 是温度参数，通常设为 1。T 的值越高，类别间的分布越平滑。

在我们的工作中，将教师网络预测的第 i 个候选区域类别的可能性记为$\{p_{i,l}^t, l \in L\}$，l 是第 l 个类别的概率，L 是类别的数目。$p_{i,l}^t$ 作为软目标从教师网络转移到学生网络，通过优化下面作为我们类别蒸馏损失的交叉熵损失 L_{cls_dis}：

$$L_{cls\_dis} = -\frac{1}{N}\sum_{i\in N}\sum_{l\in L} p_{i,l}^t log(p_{i,l}^s)$$

$P_{i,l}^s$ 是学生网络预测的类别概率，N 是候选区域的数目。软目标包含教师网络发现的不同类别间的关系信息。通过对软目标的学习，学生网络继承了这样的隐藏信息。

对调整候选区域位置和大小的边界框回归，我们使用如下 L2 优化作为边界框蒸馏损失：

$$L_{bbox\_dis} = \frac{1}{2N}\sum_{i\in N}\|R_i^t - R_i^s\|_2^2$$

Rt 和 Rs 分别表示相应的教师网络和学生网络的回归层输出。L_{bbox_dis} 可以鼓励学生网络在框回归的表现上接近教师网络性能。

掩膜的蒸馏。通过老师网络中的掩膜预测分支学生网络被训练为通道级别的蒸馏。掩膜头部的预测结果是多通道的掩码。在每个通道内，掩膜表示特定类别或简单的背景或前景的分割结果。与使用 softmax 的预测每个像素的类别的像素级别分类任务的语义分割相比，实例分割解耦了分类和分割头部来预测特定类别的掩膜。受【11】的启发，类别间的竞争不利于一个好的掩膜预测，我们设计了通道级别的掩膜蒸馏损失 L_{mask_dis} 如下：

$$L_{mask\_dis} = -\frac{1}{2N \times C} \sum_{i \in N} \sum_{c \in C} \| M_{i,c}^{t} - M_{i,c}^{s} \|_2^2$$

在我们的设定中 $M_{i,c}^{t}$ $M_{i,c}^{s}$ 是学生和老师网络在通道 c 处的候选区域 i 的 14x14 大小的掩膜输出。C 是等于类别数的通道数目。以这样的方式定义，无论类别预测如何，学生分割头部层特定类别的掩膜输出与教师网络的类似。因此，我们的语义蒸馏损失 L_{sm_dis} 是：

$$L_{sm\_dis} = L_{cls\_dis} + L_{bbox\_dis} + L_{mask\_dis}$$

这是每个头部蒸馏分割的总和。

分层蒸馏（H-Dis）

我们总体的蒸馏训练损失是分层蒸馏损失，包括特征和语义蒸馏损失，写为 L_{his_dis}:

$$L_{h\_dis} = L_{gt} + \lambda L_{sm\_dis} + \sigma L_{rp\_dis}$$

超参数 λ 和 σ 表示不同损失之间的平衡参数，在我们的实验中固定为 1。

实验

在这一节，我们对不同的主干和不同的数据集进行了评估，证明我们方法的有效性和泛化能力。具体的，我们使用带有 ResNet 的 FCIS 实例分割框架作为我们的主干。呈现在 Pascal VOC 2012 和 Cityscapes 上的结果。准确度是通过在掩膜水平的 IoU 阈为 0.5 和 0.7 的平均精度评估的。速度统计是在单张 1080Ti 显卡上对一张图片输入的所有处理过程，我们在 mxnet 实现我们的方法。

实现细节

我们遵循【17】提出的两阶段训练策略实现分层蒸馏。第一阶段是训练 RPN 网络，不需要模仿 RoIs 中的特征图。第二阶段是训练头部网络，不需要蒸馏。由于实现更高的精度不是我们的目标，我们的实现没有经过任何精度增强技巧的优化。RPN 的锚点在所有的实验中共有三个尺寸和一个长宽比。

训练：我们主要遵循 FCIS 的训练配置。我们在单 GPU 上训练所以 batch size 为 1。在 Pascal VOC 上的实验，进行 240k 次迭代，学习率分别为 10^{-3} 和 10^{-4} 在前 160K 次和后 80K 次。在 Cityscapes 上的迭代次数是 144K。在每个批次，有 128 个候选反向传播它们的梯度。所有的教师网络和学生网络是在相同的训练集和测试/验证集上进行的。

推理：在测试时，在第一个迭代时一张图片提出了 300 个候选，之后另外 300 候选在框回归分支后生成。我们的方法在测试时没有添加任何计算，所以学生网络的速度和基准模型一样。

在 Pascal VOC 上的消融实验

消融实验在 Pascal VOC 上进行。遵循【18，10，6】中的规程，对训练集执行模型训练，在验证集上进行评估，包括 20 个类的目标。训练图片被缩放为短边 600 像素。

表 1 中显示了我们工作和其它常规实例分割工作的消融实验的结果。在这些实例分割流行的工作中，我们选择 FCIS 作为我们的教师和学生实例分割架构，因为它在速度和精度上有竞争力并取得好的平衡。我们观察到，以 ResNet-18 为主干的 FCIS 框架作为我们的学生网络能达到 16fps 的高速度和与以 ResNet-101 作为我们的教师网络相比 10.2%的精度损失。我们提出的方法可以提高小的学生网络的精度到 58.1%，与 PFN 的结果接近，相比于 1fps 的 PFN，仍然可以保持 16fps 速度。

Model	$AP_{0.5}(\%)$	speed(fps)
SDS[10]	49.7	<1
CFM[5]	60.7	<1
PFN[19]	58.7	1
MNC[6]	63.5	1
Mask R-CNN[11]	69.0	3
Teacher(ResNet-101)	65.7	6
Student(ResNet-18)	55.5	16
Student w/ H-Dis	**58.1**	**16**

表 1：在 Pascal VOC 2012 数据集上，我们提出的方法和其他流行的实例分割方法比较结果。Teacher（ResNet-101）是被训练有素的教师网络。Students（ResNet-18）是没有经过蒸馏就被作为我们的基准训练的学生网络。Student w/ H-Dis 是经过我们分层蒸馏损失训练的学生网络。

表 2 展示了不同蒸馏策略，表明了不同损失的有效性。掩码头部的损失表现了最高的提升（为 AP0.5 增加了 1%）。原因是我们利用多类别的掩码预测，而不是和类别无关的掩码预测来蒸馏，其中可能包括教师网络获得的完整的特定类别信息。分类和边界框回归头部的蒸馏提升较弱，因为实例分割任务复杂且受多因素影响，仅蒸馏对数变换不能从教师网络转移丰富的知识。和离散分类的蒸馏不同，边界框回归输出可能会为学生网络提供非常错误的指导。然而，通过将损失综合起来，教师网络的知识变得有足够的影响力贡献到学生网络训练中，准确度的提升变得更有希望（为 AP0.5 总共增加了 2.6%）。

GT	Cls	Box	Mask	Mid	$AP_{0.5}$	$AP_{0.7}$
✓					55.5	36.1
✓	✓				55.8	36.1
✓	✓	✓			56.2	36.3
✓	✓	✓	✓		57.2	39.2
✓	✓	✓	✓	✓	58.1	42.3

表 2：在 Pascal VOC 2012 数据集上，进行我们的分层蒸馏时，其组件不同设置的有效性。Mid 表示中间层的特征蒸馏。

除了 ResNet-18 外，我们还在名为 ResNet-18-4 的压缩 ResNet-18 进行了实验，与原始层相比每一层的通道数减少到了 1/4。表 3 显示了 ResNet-18-4 的精度提高和速度结果，相比与基准提高了 3.5%。这个结果表明了较轻和较弱的学生网络能从训练有素的教师网络中获得更多的精度增强。我们还可以得出结论，从更好的教师网络上蒸馏能获得更大的提升，它们之间的差距越大，教师网络所传递的知识就更有效和更丰富。

Model	student	w/ H-Dis	speed(fps)
ResNet-18	55.5	58.1(+2.6)	16
ResNet-18-4	39.0	42.5(+3.5)	19

表 3：在 Pascal VOC 2012 数据集上，具有不同的主干的学生网络的比较结果。教师网络是 ResNet-101，使用 H-Dis。

为了显示我们提出的 H-Dis 方法的有效性，PSACAL VOC 数据集的实例分割结果例子显示在图 2。顶部 3 行显示，蒸馏的模型可以分割到基准结果漏掉的实例。最后三行显示我们的方法能减少一些假阳性（误检）结果。

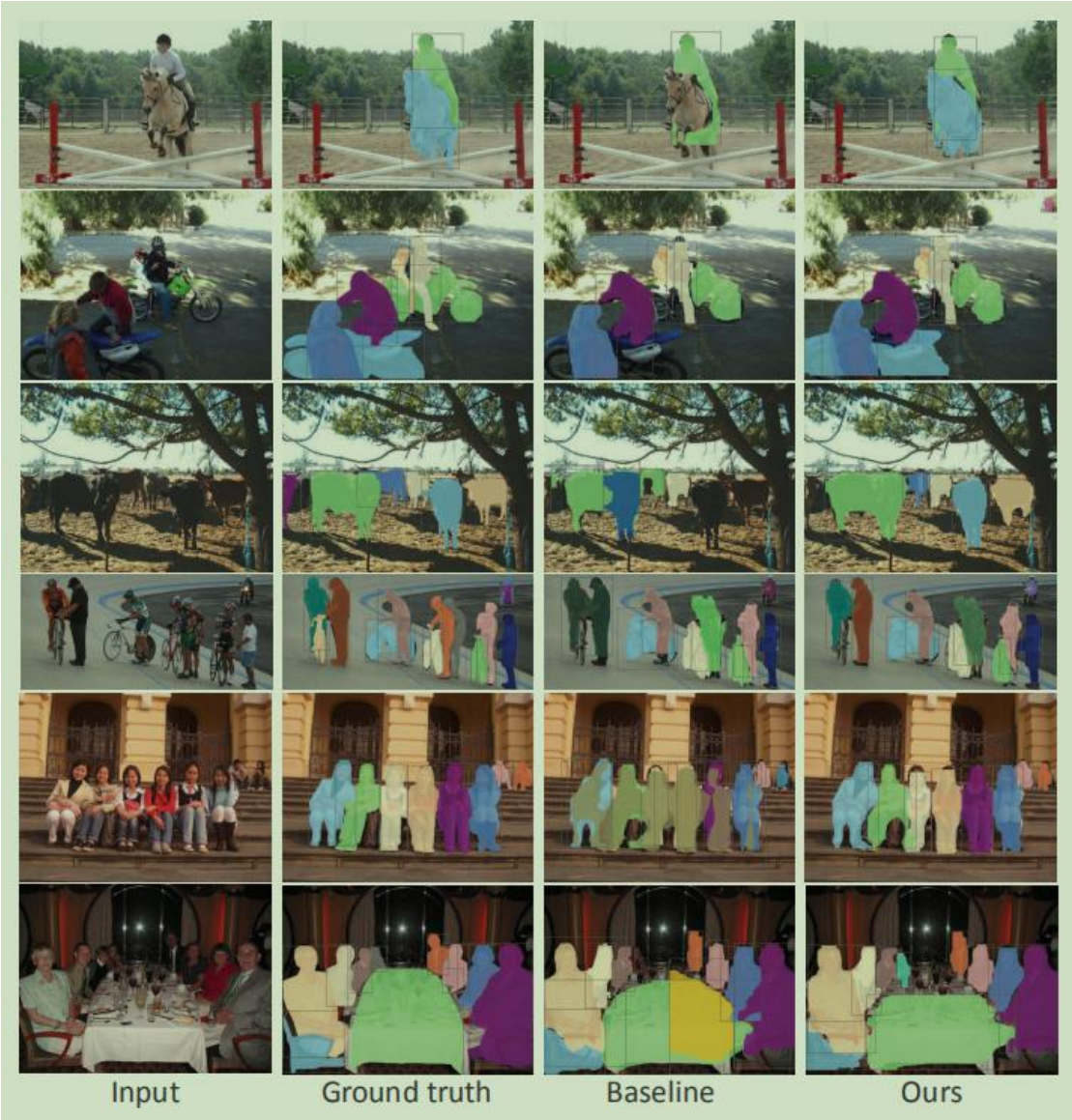

图 2：在 PSACAL VOC 2012 数据集上示例结果来自 Baseline: Student(ResNet-18) 和 Ours: Student w/ H-Dis。真实值和预测结果相对应的使用相同颜色的 mask 颜色打印。用其他颜色打印的 mask 表示假阳性（误检）结果。

Cityscapes 结果

我们进一步报告在 Cityscapes 数据集上的结果，使用 2975 张精细标注的图片训练，500 验证图片和 1525 张测试图片进行评估。图片具有 1024x2048 的高分辨率，我们在训练和测试中重新调整图片是短边为 512 像素。数据集包含 8 类用于实例分割任务，占主导地位的样本是人和车类别。

表 4 是 Cityscapes 数据集的评估结果。正如预期，学生模型在测试和验证集上分别调高了 1.6% 和 1.9%AP。在 Pascal VOC 数据集上结果提升不明显，主要是因为教师模型不够强大无法为学生模型提供有效的知识。结合表 3 中的结果，我们可以证明，教师模型没有训练足够好就来教授学生网络，提升程度会较低或几乎没有。

Model	*AP*[test]	*AP*[val]
Teacher(ResNet-101)	26.5	31.5
Student(ResNet-18)	16.5	18.6
Student w/ H-Dis	**18.1**	**20.5**

表 4：Cityscapes 数据集的比较结果。

结论

本文中，我们研究了使用分层蒸馏损失的师生学习框架来训练极小的实例分割网络的知识蒸馏。知识蒸馏通过头部和中间层来转移知识，将结构信息从笨重的教师网络蒸馏到紧凑的学生网络。在我们的语义蒸馏中除了类别和框回归蒸馏，我们引入了通道级别的掩码分支蒸馏来实现实例级别的掩码知识转移。在 Pascal VOC 和 Cityscapes 上的实验证明了我们的方法在准确性上优于基准模型，同时为实际应用提供了令人鼓舞的速度。我们对不同规模的骨干实验证明，当老师和学生的差距越大时，我们的方法越有效。