多任务知识融合策略

赵晓冬，21921062，计算机科学与技术，zhaoxiaodong@zju.edu.cn，15534733304 （**分值平分**）
季意昕，21921081，计算机科学与技术，jiyixin@zju.edu.cn，18867156831
胡单春，21921082，计算机科学与技术，3150102279@zju.edu.cn，15724998468

摘要

本文重点聚焦在深度学习模型复用上。对于不同但是较相近的视觉任务模型，我们试图将多个模型聚合，生成在功能上更加多样，但是更加轻量的模型。同时，在这一聚合过程中，不需要人工参与标定。例如，对于深度检测以及图像分割两个经典视觉问题所训练的模型，我们将其聚合后形成的轻量模型，可以应用于这两类任务，并可以扩展到其它相似视觉任务中。为此，我们提出了一种新型的训练方式，将训练目标模型参数的过程与已有教师模型紧密结合，学习得到所有教师模型的相关特征，其中误差的计算分别将其投影到对应教师模型的空间内进行计算。以此方式训练得到的模型，能够在降低模型大小的同时，仍能保证与原有教师模型相当的性能，甚至某些情况下高于教师模型。同时，生成的目标模型可以方便经过处理，应用到其它相似的视觉任务中。

关键字

知识融合，场景分析，深度估计，表面法线预测，知识蒸馏

1 引言

当下的计算机视觉领域，基于深度学习的方法在大多数视觉任务中均有绝对性的优势。但是深度学习方法仍然存在着诸多限制因素。一方面，深度学习方法需要大量的 GPU 计算资源，如果想要取得相对良好的效果，需要训练数日甚至数周；另一方面，现有的深度学习方法仍然大量依赖于人类的知识，如经过良好的标注样本等。

对于上述深度学习方法的一些不足，现在随着大量研究者将其预训练的模型发布到在线平台上，许多工作可以直接基于现有模型进行修改，极大降低了深度学习的研究成本。因此，对于如何复用现有的模型，值得相关学者的关注与研究。其中，知识融合[8]是一种较新型的方法，其可以从原教师模型中，提取特征，生成体积更小的学生模型。最近有工作[27]对这一方法进行改进，使得训练过程得到加速与优化。不过，知识融合方法存在一定的局限性：其生成的学生模型与教师模型类似，只能应用于相同的视觉分析任务中，不易进行拓展。

在传统的知识融合过程中，输入的所有预训练的教师模型均应用于同一视觉分析任务。本文提出了一个创新性的知识融合方法，可以基于多个不同相似任务的教师模型，训练得到更轻量且可以应用于多个视觉任务的学生模型。我们以场景分析与深度估计两个任务为例进行描述，同时展示了该方法可以经过简易扩展应用到其它的任务中。我们对新模型进行了大量对比实验，结果表明，学生模型在对应视觉任务中与教师模型相比有同等效果甚至更佳效果，同时，新生成的学生模型更加轻量。

2 相关工作

对于本文所研究的几个要点，我们整理了简要的综述。其中包括关于场景分析与深度估计这两项视觉任务的最近研究，以及关于知识蒸馏的相关工作进展。

2.1 场景分析

在场景分割领域，CNN（卷积神经网络）是最主流的应用模型，基于 CNN 而发展的诸多模型，在图像的场景分析中均有出色表现。最近的研究中，PSPNet[28]运用金字塔层级的

采样层，捕捉在不同尺度下的图像特征，RefineNet[13]采用了多路径的结构以利用不同层级的特征，FinerNet[26]采用了多个网络生成不同粒度下的分析特征图。SegNet[1]从另一方面做出了改进，其首先采用了编码-解码的结构，并在此之后附加了像素级别的分类层。除此之外，一些基于其它方法的模型在场景分析上也有较好的表现效果，如基于遮罩的网络模型[7, 17, 18]以及基于 GAN（生成对抗网络）的网络模型。

2.2 深度估计

传统的深度估计方法[20, 21, 22]更多采用人工提取的特征以及图模型。例如，[21]中的方法主要应用于户外场景，通过人工标定的特征，将深度估计问题转化为 MRF（马尔可夫随机场）标定问题。最近的深度估计的方法[11, 12, 14]运用 CNN 等深度学习方法，自动学习相关特征，取得了相对更加出色的效果。例如，在[4]中采用了多重尺度的深度神经网络，首先预测一个全局大致估计再对局部深度做出精细估计。

[6, 16]中的方法将深度估计与场景分割、法线估计等视觉问题相结合。PAD-Net[25]将深度估计与场景分析相结合，提出了一个多任务的预测模型以提高各自模型的表现效果。

2.3 知识蒸馏

现有的知识融合方法更多地聚焦在通过多个教师模型学习得到应用于相同任务的学生模型。训练得到的学生模型应尽可能在相同任务上保持同等效果的同时，降低模型的大小。[8]中指出经过知识蒸馏方法生成的模型在分类任务上，与教师模型具有同等的效果。[19]通过使用教师模型中间的表示结果，对原有模型进行了拓展，使其具有更深的层级但更简单的结构。[5]中的方法类似知识蒸馏，提出了一种多教师-单学生的模型进行知识蒸馏的方法，其可以应用在物体识别分类中。[23]中也从多个不同分类教师模型中学习以得到更加全面的分类模型。

除了分类领域，知识蒸馏在其它任务[2, 10, 25]中也有相应的应用。[2]中运用知识蒸馏学得一个表现更好的物体识别模型。[10]主要针对序列模型的知识蒸馏，在语音识别任务上有较好的表现。

3 预训练的教师网络

我们在此章节描述本文中涉及到的经过预训练的三个教师网络 SegNet、DepthNet 和 NormNet。我们使用这三个教师网络训练了学生模型 TargetNet。

在本文中，我们假设所有教师网络共享相同的编码器-解码器体系结构，但不限于任何特定设计，即要求所有教师网络每个环节上的输入与输出特征数相同。因为许多最先进的像素预测模型都部署了编码器-解码器架构，很容易修改这些模型参数来满足实验条件，所以本文中使用这个假设是合理的。我们将对任意架构的教师网络进行知识融合的任务作为未来工作，在本文中并不实现该目标。

适用于场景切割的教师网络（SegNet）：场景解析的目的是为图像的每个像素分配一个表示类别的标签。 在我们的实验中，我们采用具有编码器-解码器体系结构的最先进方法之一的 SegNet [1]作为场景解析任务中的教师网络。因此在场景分割中，其像素级别的损失函数可以用：

$$\mathcal{L}_{seg}(S^{gt}, S^{*})=\frac{1}{N}\sum_{i}\ell(S^{gt}, S^{*})+\lambda R(\theta), \quad (1)$$

其中，S_i^*是对像素点 i 的类别预测结果，S_i^{gt}则是真实结果，$\ell(\cdot)$是交叉熵损失，N是输入图片的像素点总数，R是 L2 范数正则项。

由于我们无法使用人工标注的标签，因此我们将教师网络的预测值S_i（S_i^*通过独热码转换而来）作为训练 TargetNet 的标注数据。

适用于深度评估的教师网络（DepthNet）：深度估计旨在为每个像素点预测一个值，该值表示目标相对于相机的深度。因此，场景解析和深度估计这两个任务之间的主要区别在于：前一个任务的输出是离散的标签，而后者的输出是连续的正数。

通过将深度值量化到N_d个桶中（每个桶长度均为ℓ），我们将深度估计任务转化为分类任务，这一方法在[15]中已经被证实是有效的。对于每个桶b，这个教师网络预测出这个目标，即这个像素点在这个桶中心的概率值，用$p(b|x(i)) = \frac{e^{r_i}}{\sum_b e^{r_{i,b}}}$计算这个概率值，其中$r_{i,b}$是网络对于像素点 i 在桶 b 的响应值。根据这个概率，连续的深度值D_i可以用以下公式进行计算：

$$D_i = \sum_b^{N_d} b \times l \times p(b|x(i)), \tag{2}$$

深度评估任务的损失函数定义为：

$$\mathcal{L}_{depth}(D^{gt}, D) = \frac{1}{N}\sum_i (d_i)^2 - \frac{1}{2N^2}\left(\sum_i d_i\right)^2, \tag{3}$$

其中$d = D^{gt} - D$，N为合法像素点的个数（我们忽略真实值缺失的像素点）。教师网络 DepthNet 的预测值D被认为是学生网络 TargetNet 的标注数据。

适用于表面法线预测的教师网络（NormNet）：给定输入图像，表面法线预测的目标是为每个像素估计表面法线向量(x, y, z)。用来训练教师网络 NormNet 的损失函数定义为：

$$\mathcal{L}_{norm}(M^{gt}, M) = -\frac{1}{N}\sum_i M_i \cdot M_i^{gt} = -\frac{1}{N} M \cdot M^{gt}, \tag{4}$$

其中，M和M^{gt}分别是 NormNet 的预测结果和真实值，M_i和M_i^{gt}分别是在像素点 i 上 NormNet 的预测结果与真实值。

4 实现方法

在本节中，我们描述了如何学习获得一个紧凑的学生网络。 我们引入了一种新颖的策略来训练学生网络，在这个方法中学生网络与教师网络一起被训练。本文提出的方法的核心是如图 1 所示的逐块学习方法。该方法通过将学生网络中蒸馏出来的知识 “投影”到每个教师网络的专业知识领域以计算损失并更新参数来学习学生网络的参数，如图 2 所示。之后，我们从 SegNet 和 DepthNet 的合并开始，然后扩展到包括表面法线预测网络（NormNet）的多个网络的合并。

4.1 学习两个教师网络

给定两个经过预训练的教师网络，SegNet 和 DepthNet 在章节 3 中进行了描述。 在图 2 中，我们训练了一个紧凑型学生网络。这个学生网络同教师网络拥有类似的编码器-解码器架构，不同之处在于解码器部分最终包括两个流向，如图 1 所示每个流解决一个任务。在本文的实现中，我们选择了某一种 VGG 架构，如[1]所述。

对于 TargetNet 的每个块，我们认为学生网络的特征图的大小要与教师网络对应块中的特征图的大小相同。TargetNet 的编码器和解码器在场景分割和深度估计的联合任务中扮演不同的角色。 编码器部分用作特征提取器，以导出两个任务的判别特征。另一方面，期望

解码器将所学习的特征“投影”或转移到每个任务域中，以便可以针对特定任务流以不同的方式使用这些特征。

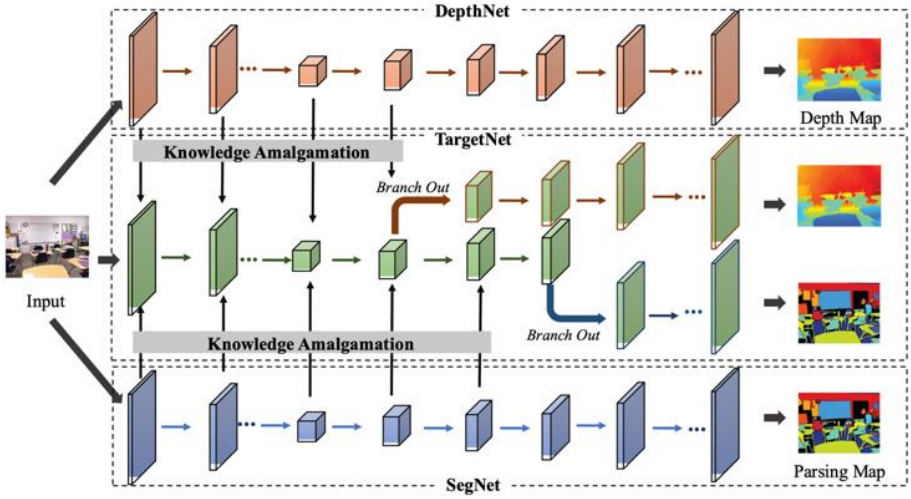

图 1. 针对场景分割和深度估计问题提出的知识融合方法。在分支操作后将替换成教师网络的解码部分

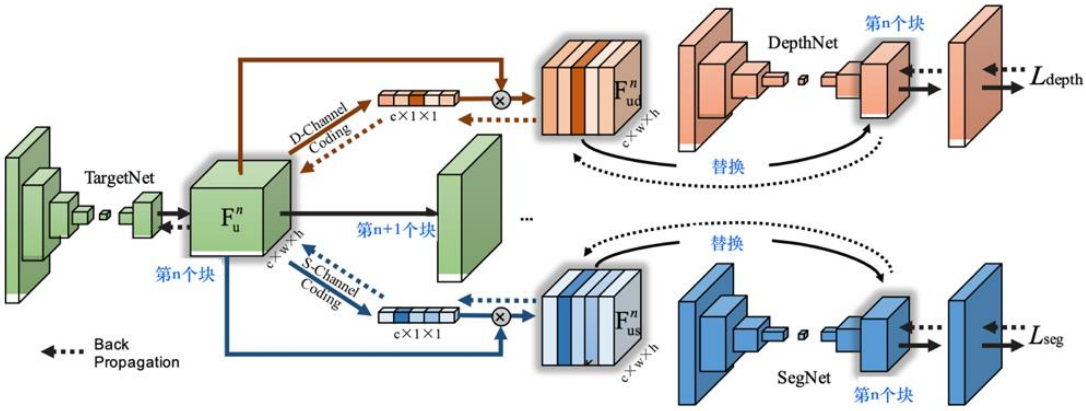

图 2. 知识融合在 TargetNet 的第 n 个块中的操作流程

尽管 TargetNet 最终有两个输出流，但使用与教师相同的编-解码器体系结构对其进行初始化。然后，我们训练 TargetNet，最后将这两个任务的两个流分支出来，然后删除分支相对应块之后的初始解码器块。 总体训练过程总结如下：

- 步骤 1：使用与教师相同的体系结构初始化 TargetNet，如 4.1.1 所述。
- 步骤 2：训练与教师网络交织在一起的 TargetNet 的每个块，如图 2 所示，详情在 4.1.2 进行阐述。
- 步骤 3：决定在 TargetNet 哪个块分支，详情在 4.1.3 阐述。

- 步骤 4：从教师网络中选择相应的块作为学生网络的分支块；删除发生在分支的块之后的所有初始块；微调 Target-Net。

在下文中，我们描述了学生网络的体系结构、损失函数和训练策略以及分支选择策略。

4.1.1 TargetNet 架构

TargetNet 用与教师相同的编码器-解码器体系结构进行初始化，其中编码器和解码器的结构是对称的，如图 1 所示。

我们采用知识融合，对学生网络的每个块，利用教师网络进行参数学习训练。记$\mathrm{F}_{\mathrm{u}}^{\mathrm{n}}$为 TargetNet 第 n 个块所提取出的特征。我们希望$\mathrm{F}_{\mathrm{u}}^{\mathrm{n}}$能同时对从教师网络中获得的场景分割和深度估计信息进行编码。为了$\mathrm{F}_{\mathrm{u}}^{\mathrm{n}}$使与教师网络互动并进行更新，我们引入了两个通道编码分支，分别称为 D 通道编码和 S 通道编码，分别用于深度估计和场景解析，如图 2 所示。直观上，可以将$\mathrm{F}_{\mathrm{u}}^{\mathrm{n}}$视为整个特征集合的容器，并且可以通过两个渠道将其投影或转换到两个任务域。

两个通道编码的架构由[9]的频道注意力模块修改而成，它由一个全局池化层和两个全连接层组成，并且大小非常轻。实际上，它仅会使参数总数增加不到 4%，从而导致非常低的计算成本。

4.1.2 损失函数与训练策略

为了在 TargetNet 的第 n 个块中学习特征$\mathrm{F}_{\mathrm{u}}^{\mathrm{n}}$，我们将未标记的样本同时输入到教师网络和 TargetNet 中，这样我们就可以获取两名教师网络的特征和 Target 的初始特征。令$\mathrm{F}_{\mathrm{d}}^{\mathrm{n}}$，$\mathrm{F}_{\mathrm{s}}^{\mathrm{n}}$表示在教师网络 DepthNet 和 SegNet 的第 n 个块上得到的特征。

对于块 n 中的蒸馏，我们首先通过将蒸馏后的特征$\mathrm{F}_{\mathrm{u}}^{\mathrm{n}}$通过 S 通道编码来获得$\mathrm{F}_{\mathrm{us}}^{\mathrm{n}}$。然后，我们用$\mathrm{F}_{\mathrm{us}}^{\mathrm{n}}$替换 SegNet 的第 n 个块的特征$\mathrm{F}_{\mathrm{s}}^{\mathrm{n}}$，并以$\mathrm{F}_{\mathrm{us}}^{\mathrm{n}}$为特征从 SegNet 获得其预测结果$\hat{S}$。以相同的方式重复此过程，以$\mathrm{F}_{\mathrm{ud}}^{\mathrm{n}}$为特征从 DepthNet 获得预测深度图$\widehat{D}$。

以这种方式，我们可以将损失函数只以$\hat{S}$、$\widehat{D}$与根据原来教师网络参数所预测出来的结果S、D来计算，公式如下：

$$\mathcal{L}_u=\lambda_1\mathcal{L}_{depth}(D,\widehat{D})+\lambda_2\mathcal{L}_{seg}(S,\hat{S}), \tag{5}$$

其中λ_1和λ_2在训练时对于 TargetNet 中所有的块都是固定的，并且$\mathcal{L}_{depth}(\cdot)$和$\mathcal{L}_{seg}(\cdot)$分别是公式 3 和公式 1。

4.1.3 分支策略

由于场景解析和深度估计是两个紧密相关的任务，因此决定在何处将 TargetNet 分支到单独的特定任务的流中以达到同时在两个任务上实现最佳性能是不容易的。与选择在编码器和解码器边界进行分支的传统多分支模型不同，我们选择了另一种方法，并发现它更有效。当使用公式 5 对 TargetNet 的 N 个块进行训练后，我们可以得到每个块相应的损失$\{\mathcal{L}_{depth}^1,\mathcal{L}_{depth}^2,\dots,\mathcal{L}_{depth}^N\}$和$\{\mathcal{L}_{seg}^1,\mathcal{L}_{seg}^2,\dots,\mathcal{L}_{seg}^N\}$。记分别对于场景分析与深度估计任务的分支块位置$p_{seg}$和$p_{depth}$，其决定方式为：

$$\begin{aligned} p_{seg} &= arg\min_n \mathcal{L}_{seg}^n \\ p_{depth} &= arg\min_n \mathcal{L}_{depth}^n, \end{aligned} \tag{6}$$

其中n满足$\frac{N}{2}<n\le N$，以保证分支块只会从解码器中选择。

一旦确定了分支块p_{seg}和p_{depth}，我们将删除最后一个分支块之后的那些初始解码器块。在图 1 所示的示例中，场景分析的分支晚于深度估计，因此将p_{seg}之后的所有初始块都删除了。然后，我们将教师网络中的相应解码器块作为 TargetNet 的分支块，如图 1 的上部和下部绿色流所示。这样，我们得到了共享一些块的、拥有一个编码器和两个解码器的 TargetNet 最终体系结构，并对该模型进行了微调。

4.2 学习更多的教师网络

我们在这里展示了两种方法，一种离线方法和一种在线方法，从三个教师网络中学习知识。我们以表面法线预测（另一个广泛研究的像素预测任务）为例来展示对三个教师网络进行知识融合。本文中所提议的两种方法可以直接应用于对任意数量的教师网络进行知识融合，只要满足之前的假设。

离线方法很简单，继续应用 4.1.2 节中的逐块学习策略。但现在我们拥有三个通道编码，其中第三个通道 M 通道编码将法线预测任务的知识进行转换。此时损失函数定义为：

$$\mathcal{L}_{u3}=\lambda_1\mathcal{L}_{depth}\left(D,\widehat{D}\right)+\lambda_2\mathcal{L}_{seg}\left(S,\hat{S}\right)+\lambda_3\mathcal{L}_{norm}\left(M,\widehat{M}\right), \tag{7}$$

其中λ_1、λ_2和λ_3是平衡权重。

在线方法以增量方式工作。假设我们已经训练了 TargetNet 用于场景解析和深度估计，为清楚起见，我们现在将原先的 TargetNet 称为 TargetNet-2。我们也希望学生网络也能融合法线预测知识，我们将完成这三项任务的学生网络称为 TargetNet-3。此在线方法的核心思想就是将原先训练所得的 TargetNet-2 作为其中的教师网络，而预训练的法线预测网络 NormNet 作为另一个教师网络，然后按照 4.1.2 节中描述的方法再次训练 TargetNet-3。

在线方法的损失函数可表示为：

$$\mathcal{L}_{u3}=\lambda_1\mathcal{L}_{norm}(M,\widehat{M})+\lambda_2\mathcal{L}_{u2}, \tag{8}$$

其中λ_1和λ_2是权重。如图 3 所示，当 TargetNet-2 作为教师网络时，在训练过程中我们将其参数固定。因此，在训练 TargetNet-3 时，我们仅更新 U 通道编码的参数，而不更新 TargetNet-2。

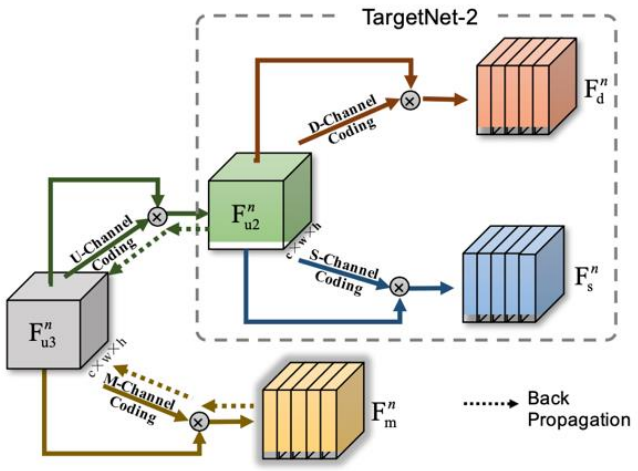

图 3. 使用 TargetNet-2 和 Norm-Net 进行知识融合。 虚线框中的部分来自 TargetNet-2，在训练 TargetNet-3 期间保持不变。

5 实验结果与分析

5.1 实验设置

数据集 NYUDv2 数据集[24]提供了 1,449 个带标记的室内场景 RGB 图像，其中包括语义分割和 Kinect 深度。还提供了 407,024 个未标记的图像，其中 10,000 个用于训练 TargetNet。除了场景解析和深度估计任务，我们还训练了 TargetNet-3，对该数据集进行表面法线预测。其中表面法线使用深度图中相邻像素的向量积计算得到。

度量指标 为了定量评估深度估计性能，我们使用相对误差的绝对值（abs rel），相对误差的平方（sqr rel）和给定阈值t，比值r在阈值1.25，1.25^2，1.25^3内的百分比作为度量标准[3]。

$$r_i = max\left(\frac{\mathrm{D_i}}{\mathrm{D_i^{gt}}}, \frac{\mathrm{D_i^{gt}}}{\mathrm{D_i}}\right)$$

其中D_i，D_i^{gt}分别为像素i处的预测值和真值。度量标准前两者越小，后者越大，性能越好。

为了评估场景语义分割性能，我们使用像素准确率（Pixel Acc.）和平均交并比（mIOU）作为指标。Pixel Acc.为正确分类的像素占所有像素的比例。mIOU 计算真值和预测值两个集合的交集和并集的比值，即真正比真正、假负、假正之和。两者越小性能越好。

为了评估表面法线预测性能，对所有有效像素点计算真值和预测值的矢量夹角，使用平均角距离，中位角距离和角距离落在三个阈值$11.25°$，$22.5°$，$30°$内的有效像素百分比作为指标[3]。前两者越小，后者越大，性能越好。

模型大小与参数数量正相关。学生网络参数数量应当小于教师网络模型的参数数量之和。

5.2 实验结果分析

表 1 中为在 NYUDv2 数据集上教师网络（TeacherNet）的结果和在解码器不同位置进行分支的学生网络的结果。TeacherNet 包含 SegNet 和 DepthNet，Decoder_bn 表示在 Decoder 的第 n 处分支出的 TargetNet。学生网络在块 4 处分出的深度估计分支和在块 5 处分出场景解析分支时，此时两个任务的性能各自最优。可以看出，学生网络在所有评测指标上均优于教师网络，这表明我们所提出的方法是有效的。网络规模方面，从 Decoder_b[1-5]的参数数量可以看出，合并的特征块越多，目标网络越小。也就是说，解码器的分支更靠后将会得到更小的网络。最终的学生网络分别在第 4 块和第 5 块进行分支，得到 Target-D 和 Target-P，用于深度和场景，网络规模大小约为教师网络的一半。

表 1. 在 NYUDv2 数据集上教师网络和具有不同分支的学生性能比较

方法	参数	mIOU	Pixel Acc.	abs rel	sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
TeacherNet	~55.6M	0.448	0.684	0.287	0.339	0.569	0.845	0.948
Decoder_b1	~36.9M	0.447	0.684	0.276	0.312	0.383	0.753	0.939
Decoder_b2	~31.0M	0.451	0.684	0.259	0.275	0.448	0.799	0.939
Decoder_b3	~28.3M	0.451	0.684	0.260	0.277	0.448	0.796	0.951
Decoder_b4 (Target-D)	~28.0M	0.452	0.683	**0.252**	**0.257**	**0.544**	**0.847**	**0.959**
Decoder_b5 (Target-P)	~27.8M	**0.458**	**0.687**	0.256	0.266	0.459	0.810	0.956

表 2 中为以离线方式训练的 Target-3 和 Target-2、TeacherNet 的结果比较。可以发现，TargetNet-3 中的表面法线估计任务帮助提高了场景分割和深度估计的精度，超过了 TargetNet-2。另外，由于表面法线与深度估计任务关系更加紧密，深度估计的性能增加更为显著。

表 2. 在 NYUDv2 数据集上的 TeacherNet（SegNet，DepthNet 和 NormNet）,Target-2 和 Target-3 在线训练的性能比较

方法	mIOU	相对差		角距离		阈值		
		abs	sqr	Mean	Median	11.25°	22.5°	30°
TeacherNet	0.448	0.287	0.339	37.88	36.96	0.236	**0.450**	0.567
TargetNet-2	0.458	0.252	0.257	-	-	-	-	-
TargetNet-3	**0.459**	**0.243**	**0.255**	**35.45**	**34.88**	**0.237**	0.448	**0.585**

表 3 中为以在线方式训练的 TargetNet-3 和 TargetNet-2、TeacherNet 的结果比较。与表 3 中不同，此处 TargetNet-3 以 NormNet 和 Target-2 为教师网络进行训练。可以发现，TargetNet-3 在场景解析和深度估计方面的性能都比 TargetNet-2 更好，在表面法线预测方面优于 NormNet。

表 3. 在 NYUDv2 数据集上的教师网络，TargetNet-2 和在线训练的 TargetNet-3 的比较结果

方法		参数	场景分析		深度估计		表面法线	
			mIOU	PA	abs rel	sqr rel	mean angle	median angle
Teacher for TargetNet-2	SegNet	~83.4M	0.448	0.684	-	-	-	-
	DepthNet		-	-	0.339	0.287	-	-
Teacher for TargetNet-3	NormNet		-	-	-	-	37.88	36.96
	TargetNet-2	~27.8M	**0.458**	0.687	0.256	0.266	-	-
TargetNet-3	Decoder_b1	~46.1M	0.452	0.680	0.258	0.269	37.4	32.7
	Decoder_b2	~34.2M	0.455	0.687	0.253`	0.254	36.9	32.0
	Decoder_b3	~28.8M	0.457	0.687	**0.253**	**0.253**	36.3	31.4
	Decoder_b4	~28.1M	**0.458**	**0.688**	0.258	0.271	36.1	31.0
	Decoder_b5	~27.9M	0.457	0.683	0.261	0.278	**35.5**	**30.3**

6 结论

在本文中，我们提出了一种新的知识融合策略，用于学习多个不同任务的教师，在不需要人工标注数据的前提下得到多功能学生模型。首先通过两个教师网络训练学生模型，合并不同教师的特征。接着将该方法推广到三个及多个教师网络的蒸馏中，并提供了离线训练和在线训练两种方式。在多个大型数据集上的测试结果表明，学生网络能够处理所有教师的任务，性能与教师媲美，同时还具有轻量级的特性。在未来的工作中，我们将会探索结构更复杂的教师模型的知识融合，如 ResNet，DenseNet 等；以及不同结构的教师模型融合，尝试弥合教师特征图之间的语义差异。

参考文献

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2481–2495, 2017.

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017.

[3] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convo- lutional architecture. *International Conference on Computer Vision*, pages 2650–2658, 2015.

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depthmap prediction from a single image using a multi-scale deep net- work. *Neural Information Processing Systems*, pages 2366– 2374, 2014.

[5] Jiyang Gao, Zijian Guo, Zhen Li, and Ram Nevatia. Knowledge concentration: Learning 100k object classifiers in a single cnn. *arXiv: Computer Vision and Pattern Recognition*, 2017.
[6] Jean-Yves Guillemaut and Adrian Hilton. Space-time joint multi-layer segmentation and depth estimation. In *3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 440–447, 2012.
[7] Kaiming He, Georgia Gkioxari, Piotr Dolla´r, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, pages 2980–2988, 2017.
[8] Geoffrey E Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *Neural Information Processing Systems*, 2015.
[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Computer Vision and Pattern Recognition*, 2018.
[10] Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu. Knowledge distillation for sequence model. *Proc. Interspeech 2018*, pages 3703–3707, 2018.
[11] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *International Conference on 3d Vision*, pages 239–248, 2016.
[12] Bo Li, Yuchao Dai, and Mingyi He. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognition*, 83:328–339, 2018.
[13] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *Computer Vision and Pattern Recognition*, pages 5168–5177, 2017.
[14] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. *Computer Vision and Pattern Recognition*, pages 5162– 5170, 2015.
[15] Arsalan Mousavian, Hamed Pirsiavash, and Jana Kosecka. Joint semantic segmentation and depth estimation with deep convolutional networks. *International Conference on 3d Vision*, pages 611–619, 2016.
[16] Haesol Park and Kyoung MuLee. Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence. *International Conference on Computer Vision*, pages 4623–4631, 2017.
[17] Pedro H O Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. *Neural Information Processing Systems*, pages 1990–1998, 2015.
[18] Pedro H O Pinheiro, Tsungyi Lin, Ronan Collobert, and Pi- otr Dollar. Learning to refine object segments. *European Conference on Computer Vision*, pages 75–91, 2016.
[19] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *International Conference on Learn- ing Representations*, 2015.
[20] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2006.
[21] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Learning 3-d scene structure from a single still image. In *International Conference on Computer Vision*, pages 1–8, 2007.
[22] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. *Computer Vision and Pattern Recognition*, 1:195–202, 2003.
[23] Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating knowledge towards comprehensive classification. In AAAI Conference on Artificial Intelligence, 2019.
[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012.
[25] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. Computer Vision and Pattern Recognition, pages 675–684, 2018.
[26] JingwenYe,ZunleiFeng,YongchengJing,andMingliSong. Finer-net: Cascaded human parsing with hierarchical granularity. In International Conference on Multimedia and Expo, pages 1–6, 2018.
[27] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Computer Vision and Pattern Recognition, 2017.

[28] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In Computer Vision and Pattern Recognition, pages 2881– 2890, 2017.