不平衡数据下用于机器人感知的多任务学习

Özgür Erkent[1]

[1]哈杰泰佩大学　　论文，工程学院，计算机工程　　系，安卡拉

提交日期 / 收到日期：xx.xx.202x 接收日期 / 录用日期：xx.xx.202x

摘要

研究表明，多任务问题求解能够提升单项任务的准确率。对于资源有限的机器人而言，这一特性尤为重要。然而，当各项任务的标签数量不相等，即存在不平衡数据时，样本数量不足可能会引发问题，而且移动机器人在各种环境中获取标签并非易事。我们提出了一种方法，即使某些任务缺少真实标签，也能学习这些任务。我们还对所提出的方法进行了详细分析。一个有趣的发现与任务之间的相互作用有关。我们展示了一种方法，用于找出哪些任务能够提升其他任务的性能。我们通过将任务输出（如深度）作为输入来训练教师网络，对此进行研究。我们还提供了使用少量数据进行训练时的实验证据。我们在不同的数据集 NYUDv2 和 Cityscapes 上使用了语义分割和深度估计任务。

关键词：自主机器人；分割；深度估计；多任务学习中的不平衡数据

面向不平衡数据的机器人感知多任务学习

Öz

已观察到，在机器人中解决多任务问题能够提升单项任务的性能。由于机器人资源有限，这一特性十分重要。然而，当各项任务的标签数量不相等，即存在不平衡数据时，会因样本数量不足而产生问题，并且在不同环境中为移动机器人进行标注并不容易。在本研究中，我们提出一种方法，即使某些任务缺少真值标签，也能够学习全部任务。我们还对所提出的方法进行了详细分析。一个有趣的发现与任务之间的相互作用有关。我们展示了一种方法，用以找出哪些任务能够提升其他任务的性能。我们通过在教师网络的训练中将深度等其他任务的输出作为输入来对此进行研究。此外，我们还提供了在少量数据上训练时的实验证据。我们在NYUDv2和Cityscapes等不同数据集上使用语义分割和深度估计任务。

关键词 关键词：自主机器人；分割；深度估计；多任务学习中的不平衡数据

通讯作者： Özgür Erkent ozgurerkent@hacettepe.edu.tr

引言

自主机器人需要完成多种感知任务，才能理解周围环境并据此采取行动。近年来，能够同时学习多项任务的深度神经网络（DNN）已成为重要的研究课题（Kokkinos，2017年；Crawshaw，2020年；Vandenhende 等，2022）。研究表明，多任务学习（MTL）方法的性能可接近单任务方法；此外，在某些条件下，其性能甚至可以超越单任务方法（Kendall、Gal 和 Cipolla，2018年；Sener 和 Koltun，2018年；Zamir 等，2018年；Standley 等，2020年）。然而，多任务学习主要依赖带标签数据进行训练，而为所有任务获取带标签数据十分繁琐。在某些情况下，可能无法获得所有任务的标签。当数据集中各类别的表示数量不相等时，该数据集便称为不平衡数据集（Chawla、Bowyer、Hall 和 Kegelmeyer，2002）。在多任务学习场景下，只有少数研究关注了这一问题（Ghiasi、Zoph、Cubuk、Le 和 Lin，2021年；Xu、Yang 和 Zhang，2023）；但这些研究主要着眼于为缺失任务标签训练一个教师网络，并为每项任务使用单模态分类器，却没有利用标签过多的任务所提供的额外标签信息。

在本研究中，我们提出了一种方法，通过使用多模态教师网络，找出哪些任务能够提升其他任务的性能。我们将其中一个任务可用的真实标签与原始传感器输入一并输入教师网络，以生成另一个任务缺失的标签。随后将这些标签作为伪标签，用于训练学生网络。我们使用多模态网络作为教师，因为它能够探索更多信息，为学生网络生成伪标签，从而提高准确率。值得注意的是，文献中此前已提出单模态方法（Ghiasi, Zoph, Cubuk, Le, & Lin, 2021）；然而，这些方法无法利用所有可用数据。我们分析了假设成立的条件，因为众所周知，多任务学习并非在所有条件下都能适用于所有任务（Standley, et al., 2020）、（Vandenhende, et al., 2022）、（Milli, Erkent, & Yılmaz, 2023）。与其他研究不同，其他研究需要大量平衡数据来找出相互支持的任务，而我们开发了一种方法，用于分析能够相互提升性能的标签

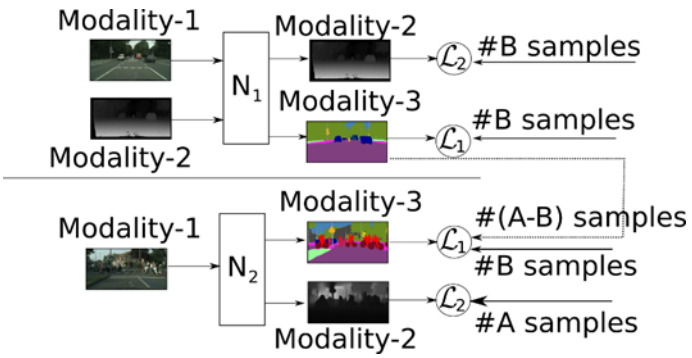


Figure SEQ Figure * ARABIC 1. A sample multi-modal multi-task teacher DNN at the top () and a uni-modal multi-task student DNN at the bottom with two tasks (). Modality-1 is RGB image, Modality-2 (Mod2) is depth and Modality-3 (Mod3) is semantic segmentation in this example. and compute the losses for the tasks. We assume we have #A labeled samples for Mod2 and #B labeled samples for Mod3 where #A>#B for this example. First, the top DNN optimizes its parameters by using the #B samples. Then, #(A-B) pseudo labels are obtained by infering the top DNN and these labels are provided to compute the loss function of the ; therefore uses all samples efficiently to optimize its parameters

样本数量少且分布不均衡的多任务学习设置。

我们的贡献如下：

- 我们展示了一种方法，在训练过程中将其中一项任务用作输入模态，以找出能够提升其他任务性能的任务。

● 我们分析了多模态教师网络生成的多任务学习中的伪标签对高效优化参数的影响，

● 我们针对多个数据集，为在缺失标签的不平衡数据上使用伪标签的影响提供了实验证据，

我们提出的方法概览如图1所示。此处，使用 B 个N_1样本进行训练，而则使用一定数量的样本进行训练，其中包括从N_2网络获得的伪标签。我们在方法部分分析其效率，并在实验N_1部分展示其性能。我们将首先在相关工作中讨论相关研究。最后，我们通过简短的结论总结我们的工作。

相关工作

由于多任务学习能够提升性能，因此自其提出以来一直在不同设置下受到研究，这种提升既体现在准确率方面，也体现在计算时间方面。有关详细综述，请读者参阅该主题的最新综述（Crawshaw，2020年）、（Vandenhende 等，2022年）。一项早期研究表明，多任务深度神经网络 UberNet 能够同时执行七项不同任务，且各项任务的性能不会下降（Kokkinos，2017年）。

随着多任务网络提升了某些任务的性能，人们收集了包含多种任务的数据集，例如 Taskonomy（Zamir 等，2018年）。然而，研究表明，并非所有任务都能相互支持（Standley 等，2020年）。这是一个关键且重要的实证发现。关于如何确定哪些任务能够相互支持，目前尚未提供详细的理论背景。在本研究中，我们提出了一种方法，无需进行大量训练，也无需使用大型带标签数据集，即可找出哪些任务能够相互提升性能。

根据使用方式的不同，人们针对多任务网络开发了不同的解决方案。一些方法使用切换网络，在每次选择一个特定任务；尽管这些方法能够达到较高的性能（Sun 等，2021年），但其计算时间和内存需求通常超出了移动机器人平台的承受范围。我们将考虑使用共享权重层（通常为共享骨干网络）并同时执行多个任务的网络，因为共享骨干网络在计算时间和内存需求方面具有重要优势。例如，（Rottmann、Posewsky、Milioto、Stachniss 和 Behley，2021年）使用初始多任务网络来提升深度估计的性能，然后仅使用深度标签训练单任务网络。（Lee、Rameau、Im 和 So，2022年）利用语义先验从单目摄像机图像中估计深度。（He、Yang、Zhang、Long 和 Song，2024年）从任务难度的角度考虑任务平衡。PAD-Net（Xu、Ouyang、Wang 和 Sebe，2018年）将初始估计作为先验，然后通过多模态蒸馏模块对其进行融合。一项后续工作 MTI-NET（Vandenhende、Georgoulis 和 Van Gool，2020年）再次计算初始估计；不过，这一次是在不同尺度上进行计算，并通过特征传播模块融合来自这些尺度的特征，最后再次使用多模态蒸馏模块融合这些特征。尽管这些方法在多任务学习设置下能够在深度估计和语义分割方面取得较高的准确率，但它们假设每个任务都拥有平衡的数据集。然而，在数据集中为所有样本添加标签并不总是可行，尤其是在移动机器人平台上，这会导致不平衡数据。需要注意的是，在机器人研究中，获取深度信息通常比获取语义分割标签容易得多。

尽管类别不平衡数据问题不一定局限于多任务学习场景，但此前的研究已经对其进行了探讨。为解决数据不平衡问题，研究者采用了数据级平衡方法（Chawla、Bowyer、Hall和Kegelmeyer，2002）；然而，这些方法无法利用所有可用数据，因此可能会丢失部分标注信息。损失加权方法会为样本数量较少的类别分配较大的权重。这种方法也用于多任务学习任务，例如为不存在对应任务样本的损失赋予零权重（Leang、Sistu、Burger、Bursuc和Yogamani，2020）；然而，正如实验结果所示，这并不总能提高两个任务的准确率。最后，我们将介绍用于解决类别不平衡数据问题的生成对抗网络（生成对抗网络）：它通过生成对抗网络生成缺失样本；但当样本数量不足以生成这些合成样本时，问题就会出现（Wang等，2022；Choi、Jung、Kim和Yoon，2022）。我们的方法利用可用样本生成伪标签，因此不需要合成样本；此外，我们还研究了样本不足以生成合成数据的情况。

在实验中将会看到，这并不总能提高两个任务的准确率。最后，我们将介绍用于解决类别不平衡数据问题的生成对抗网络（生成对抗网络）：它通过生成对抗网络生成缺失样本；但当样本数量不足以生成这些合成样本时，问题就会出现（Wang等，2022；Choi、Jung、Kim和Yoon，2022）。我们的方法利用可用样本生成伪标签，因此不需要合成样本；此外，我们还研究了样本不足以生成合成数据的情况。

少数研究通过采用教师-学生网络来解决这一问题。例如，Ghiasi、Zoph、Cubuk、Le 和 Lin（2021年）使用单模态教师为学生模型生成伪标签。但在这种情况下，额外模态提供的额外信息并未得到有效利用。我们提出了一种解决这一问题的多模态方法，利用额外标签训练多模态网络。表 1 概述了这些方法。我们的方法是唯一一种在多任务设置中考虑不平衡数据，并利用传感器融合训练教师网络的方法。

在下一节中，我们将深入分析我们的方法，并分析三个条件。然而，由于多任务问题的上界难以求得（Boursier、Konobeev 和 Flammarion，2022年），我们在实验部分提供了实证见解。

表 1. 考虑多任务分类和不平衡数据标注的方法

	MT	不平衡数据考量	使用以下内容进行教师网络训练传感器融合
（Vandenhendeet. Al.，2020年）	+	-	-
(Chawla 等，2002)	-	+	-
(Leang 等，2020年)	+	-	-
(Wang 等，2022)	+	+	-
(Ghiasi 等，2021年)	+	+	-
(Xu、Yang 和 Zhang，2023)	+	+	-
Ours	+	+	+

方法

假设我们有两个函数，它们能够根据输入生成输出，即 f_1 f_2 z和 。我们还假设有第三个多任务函数，它可以同时输出这两个函数 $X_1 = f_1\ z() X_2 = f_2\ z()$，其中 是另一个函数，$g\ f_3\ z()\bigcirc= f_1\ z(),$ $f_2\ z()\bigcirc f_3$，该函数可以视为用于生成特征的函数（类似于深度神经网络中的层）。

我们有一组输入 的标签，以及输入 的标签。也就是说，我们没有 X_1^a, X_2^a $z^a,$ $a\in A$ X_1^b $z^b, b\in B$ 的标签。我们将提供估计函数；并利用可用数据 X_2^b $\hat{g}\ f_3\ z(\ );\theta_g()\hat{f}_3\ z;\theta_3()$ 和标签，其中 和 是待优化的参数。 θ_g θ_3

首先，我们可以将函数 、 和 近似表示为其子函数之和：$f_1\ f_2$ f_3

$$\widetilde{f}_a(z) \approx f_{a0}(z) + f_{a1}(z) + f_{a2}(z) + \cdots = \sum_{n=0}^{N} f_{an}(z) \qquad (1)$$

注意，。我们的主要假设是，的子函数可以表示为的子函数的$f_a - \widetilde{f}_a$ $()$ $= 0$ f_3线性组合，如下所示：f_1 f_2

$$f_3(z) \approx \sum_{i\in I} f_{3i}(z) + \sum_{j\in J} f_{3j}(z) + \sum_{k\in K} f_{3k}(z)$$

$$f_{3i}(z) \approx \alpha_l f_{1l}(z) \approx \beta_m f_{2m}(z);\ l \in L,\ m \in M \quad (2)$$
$$f_{3j}(z) \approx \alpha_o f_{1o}(z);\ o \in O f_{3k}(z) \approx \beta_p f_{2p}(z);\ p \in P$$

其中，和是超参数；是的子函数集；是$\alpha > 0$ $\beta > 0$ $I,\ J,$ K f_3 $L,\ O$的子函数集；而是的子函数集。也就是说，如果正在生成特征，那么它的f_1 $M,\ P$ f_2 f_3特征可以利用函数（的子函数）和的特征生成。f_1 f_2

我们可以使用深度神经网络来近似 和 ，其中 g f_3 $\hat{g}\ f_3\ z$ (); θ_g $\hat{f}_3\ z;\theta_3$为参数；只要有足够数量的带标签数据样本 ，这些参数便可通过随机方式获得。的输出可以与 网络中获得的特征相似。同样，如果 和 函数由神经网络实现，那么它们的子函数也可以是 网络中获得的特征。f_2

我们使用多模态多任务估计器来估计标签（用于教学的 h $X_1,\ z;\theta_h$ $X_1,\ X_2$网络）。请注意，预计它在 上的一个用于特征提取的骨干网络。其参数 X_1
计器 $X_1^a,\ X_2^a$ X_2^b估计伪标签。伪标签是教师网络 的输出。h $X_1,\ z;\theta_h$ 教师网络是一标签对应的伪标签。伪标签被用作缺失的真实标签，以训练学生网络。单模态估计器将使用 $\hat{g}\ f_3\ z();\ \theta_g$ 标现了平衡。如果 和 函数共享相似的子函数（特征），那么学习一个任务的权重就可能有助于学习另一个任务的

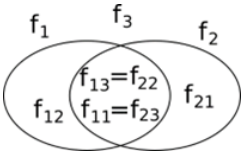


图2. 条件1在一个非常简单案例中的示例说明。

条件*1*：并且，也就是说$\|f_{3i}\| > \|f_{3j}\|$ $\|f_{3i}\| > \|f_{3k}\|$任务之间共享子函数的数量很高。那么多模态教师网络h将能够优化其用于任务估计的参数，因为它能够通过以平衡的方式使用来自和的f_2 f_1 f_2标签，同时学习用于和的特征。由于其中大多数是共同子函数（从神经网络中大型骨干网络获得的X_1^a X_2^a特征），因此我们需要学习较少数量的权重；换句话说，我们需要更少的样本来学习相同数量的~权重。那么，伪标签将具有较高的估计准确率。请注意，如果我们使用单模态教师网络，就不会X_2^b拥有这一优势。图2展示了一个非常基础的示例说明。这里，子函数被直接表示为相等；然而，它们也可以是公式 2中给出的线性组合。总共有4个子函数，而和都有3个子函数。在此
f_3 f_1 f_2

图中，由于共同子函数的数量高于非共同子函数，预计两项任务都会提升彼此的性能。我们还有另外两个$\|f_{3i}\| > \|f_{3j}\| = \|f_{3k}\|$条件需要考虑。

条件*2*：并且，的子函数数量比有更多与相似的子函数（特征），而不是。在这种情况下，将作为输入提供给 f_2 f_1 X_1，$z()$函数，即，不会提高对h h X_1，$z()$伪标签的估计性能，因为要预测的特征（子函数）会被学习，但要估计的大多数特征（子函数）不会像那样被学习。另一方面，如果我们将作为输入提供给函数，，就会改善对伪标签，$z()$h h X_2，$z()$的估计性能，因为要预测的特征（子函数）也会构成如那样要预测的特征（子函数）中的大多数。X_2 $\|f_{3i}\| > \|f_{3j}\|$

该条件也存在镜像情形；我们简要说明，对于 和，$\|f_{3i}\| > \|f_{3k}\|$ $\|f_{3j}\| > \|f_{3i}\|$那么输入将提升 估计 的性能，但输入，也就是X_1，$z( )$ h X_2 X_2，$z( )$函数，并不会提升估计 的性能。h X_2，$z( )$ X_1

因此，我们可以得出结论：如果一个任务的子函数（特征）符合这一条件，那么其中一个任务将提升另一个任务的性能，但另一个任务不会带来提升。

条件 *3*：在该条件下， 和 的函数所具有的
$\|f_{3j}\| > \|f_{3i}\|$ $\|f_{3k}\| > \|f_{3i}\|$ f_1 f_2共同子函数少于不共同的子函数。因此，多任务训练不会提高这两个函数的准确率。多模态性不会提供准确的伪标签。这可能有两个主要原因。

第一个原因可能是，所提供的样本数量不足以找到和的子函数的适当近似，并进行随机处理。在这种情况下，无法发现这两种模态之间的相似性f_1 f_2，性能也无法提升。第二个原因可能是，即使提供了足够的样本，f_1和的子函数（特征）之间也可能只有少量相似之处。f_2

实验

玩具示例

首先，为了更清楚地说明，我们使用两个简单函数来评估我们的方法，

$$f_1(z) = sin(z) + \sqrt{z}/300.0 + 2 + 3sin(5z)cos(2z)$$

$$f_2(z) = cos(z/3) + \sqrt{(z)}/200.0 + 1 - sin(3z) * cos(4z)$$

我们尝试使用一个包含f_2 4 128 层、每层含有若干神经元的全连接神经网络来近似。如果该网络是单任务（ST）网络，则只有一个输出值；如果是多任务（MT）网络，则有两个输出，分别用于估计两个函数，即和。输入变量（z）取自f_1 f_2与[0,1], 之间，其中有640个随机变量用于，有160个随机变量用于。首先，我们用160个点训练单任务网络，结果如图3所示。可以看出，结果并不理想，如表2所示，误差为。接下来，我们使用类似的0.0149 ± 0.0111

网络，并将其修改为有两个输出，一个用于，一个用于。我们使用f_1 $f_2$160个样本训练该网络，观察到结果有所改善。最后，我们训练一个教师多任务网络（网络），其输出为和，并使用f_2 f_1 =160 $\tilde{f}_2$个样本。然后，我们使用640个样本生成伪h $z,f_1()$输出。我们使用这些伪f_1 $\tilde{f}_2$值训练学生单任务网络。可以看出，这种设置的准确率最高，与多任务学习相比，准确率提高了约30%。需要指出的是，这两个函数符合条件1，结果也支持我们的假设。

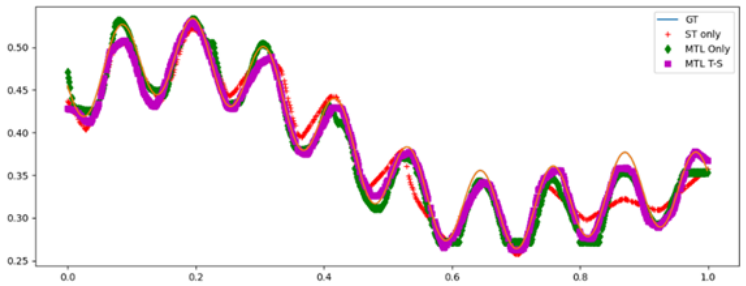


图 3. 玩具示例的结果。真实值：真实值。单任务：单任务。多任务学习 T-S：教师-学生方法的多任务学习。建议以彩色查看。

表 2. 玩具示例的平均均方误差

ST	仅多任务学习	多任务学习 T-S
0. 0149 ± 0. 0111	0 0101 ± 0 0068 . .	0. 0072 ± 0. 0062

NYUDv2

接下来，我们在 NYUDv2 数据集（Nathan Silberman & Fergus，2012）上评估多模态教师网络和单模态学生模型。该数据集由使用 RGBD 相机拍摄的室内图像组成，包含 RGB图像及其对应的深度信息，并涵盖 40 个语义分割类别。

我们采用一种通过蒸馏进行语义分割和深度估计的多任务深度神经网络（Vandenhende, Georgoulis, & Van Gool, 2020），并对其进行修改，加入用于多模态的特征融合模块（FFM），如图4所示。网络的具体结构如下。我们使用 HRNet（HRNet_w18）作为**骨干网络**，从RGB图像中提取特征（Wang, et al., 2020）；RGB部分使用在 ImageNet 上预训练的权重。如果使用深度图像或语义分割图像等额外模态，则通过另一个 HRNet_w18 提取其特征。学习率为 0.0001，批量大小为 4。请注意，只有教师网络可以使用额外模态。**FFM**通过在每个尺度层级将不同模态的特征相加来进行融合，这是我们对多模态多任务网络的贡献。

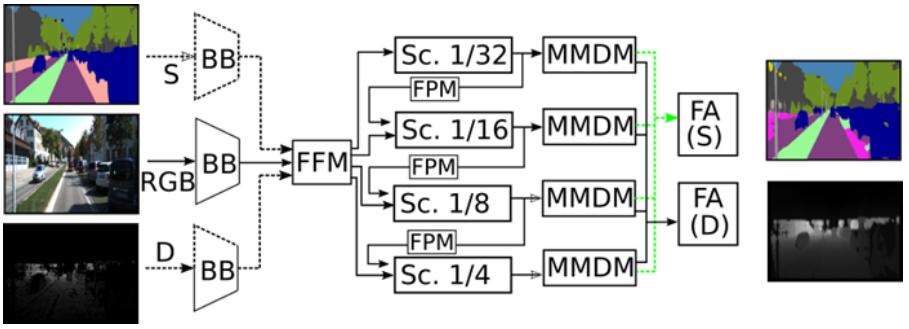


图4。 实验中使用的网络。BB：骨干网络。FFM：特征融合模块。FPM：特征传播模块。MMDM：多模态蒸馏模块。FA：特征聚合。正文简要介绍了各模块的细节。若将语义分割图像或深度/激光雷达图像作为输入，则使用黑色虚线。

我们使用特征金字塔网络（FPN）融合四种不同尺度（1/4、1/8、1/16、1/32）的特征。每种尺度下的初始估计均作为特征，并通过**FPM**传递给更大尺度的估计；该模块会对特征进行卷积和上采样。最后，在不同尺度的特征上应用**MMDM**，并由**FA**通过关注最相关的特征对其进行聚合，从而得到最终的任务预测。训练时我们使用以下总损失

$$L = L_D + \alpha L_S$$

其中，表示语义分割的交叉熵损失，表示深度估计的平均绝对损失。是用于平衡各任务损失权重的参数。我们使用了α。这一决定是在训练数据上进行交叉验证研究后作出的。由于我们想测试假设数据不平衡的方法，因此使用了不同于原始数据集的自定义数量的训练图像。我们使用500张深度图像和250张用于语义分割的已标注图像，并使用395张图像进行验证。这是机器人学中的典型情况：深度图像的数量多于语义分割图像。由于标签数量远少于原始数据集，预期结果的准确率也会低于使用完整数据集训练得到的结果。首先，我们进行了消融研究，以考察使用伪标签的效果。作为对照，输入为单张RGB图像，输出为深度和语义（单模态多任务，UMMT）。使用500个深度标签和250个分割标签训练UMMT。使用250个语义标签和深度标签训练教师网络，以生成伪标签（UMMT-PS）。随后，使用250个没有对应语义标签的深度标签来获取250个新的伪标签。将这些伪标签与原有的500个深度标签和250个语义标签一起用于训练学生模型。结果见表3。消融研究表明，伪标签方法取得了更好的结果。

表3. NYUDv2 消融研究。

	mIoU	rmse	对数均方根误差	||f1||	||f2||
UMMT	22.1	0.790	0.264	D	S
UMMT-PS	**25.3**	**0.653**	**0.206**	D	S

表 4. NYUDv2 结果。

	mIoU	Rmse	对数均方根误差	输入	Out	loss	||A||	||B||	||v||	||f1||	||f2||
MMST	24.0	-	-	RGB + D	S	GTS	250	-	395	S	S
MMMT -D	24.4	*0.337*	*0.102*	RGB + D	语义分割 + 深度	GTS + GTD	500	250	395	D	S

MMMT-S	*52.6*	0.805	0.261	RGB + S	语义分割 + 深度	GTS + GTD	500	250	395	S	D
FMST	19.7	-	-	RGBD	S	GTS	250	-	395	S	S
UMST-S	20.7	-	-	RGB	S	GTS	250	-	395	S	S
UMMT	22.1	0.790	0.264	RGB	语义分割 + 深度	GTS + GTD	250	-	395	D	S
UMMT-D	19.2	0.783	0.243	RGB	语义分割 + 深度	GTD + GTS	500	250	395	D	S
UMMT-PD	23.9	0.882	0.272	RGB	语义分割 + 深度	PSD + GTD + GTS	500	250	395	S	D
UMMT-PS	**25.3**	**0.653**	**0.206**	RGB	S + D	PSS + GTS + GTD	500	250	395	D	S

结果如表4所示。表中的缩写含义如下：D：深度，S：语义分割，GTS：S的真实值，GTD：D的真实值，PSS：S的伪标签，PSD：D的伪标签，MM：多模态，UM：单模态，ST：单任务，MT：多任务。||A|| 和|B||：深度和语义分割的已标注图像数量（包括伪标签）。所有情况均使用了295张验证图像。为与方法部分中的条件进行比较，也提供了和。平均交并比（mIoU）为 f_1 f_2 用于语义分割测量（越高越好），均方根（均方根误差）和对数均方根误差用于深度估计测量（越低越好）。

首先，我们分析多模态和多任务对分割的影响。最佳结果以粗体显示，而斜体结果表示该输出被用作输入，因此不纳入比较。

我们对所有情况使用与图4所示相同的架构。由于单任务网络中只有单一估计输出，我们在单任务中不使用任何FA模块。第一行是以RGB为输入的标准单分割任务（UMST-S）。在单模态中，仅使用RGB输入分支，不使用FFM。

接下来，我们使用多任务而非单任务，同时进行深度估计和语义分割任务（UMMT-D）。使用了500个深度标签和250个分割标签。如果语义分割对相应的深度标签没有标签，则取为零，这是多任务中的标准做法L_S

相关研究（Leang、Sistu、Burger、Bursuc和Yogamani，2020年）。然而，平均交并比性能从20.7%降至19.23%，均方根误差为0.7825。这是由于深度分支训练过度，导致语义分割性能下降。

接下来，我们使用多任务网络，并确保各任务的标签数量相同，每个任务各250个（UMMT）。平均交并比提升至22.17%，均方根误差为0.7896。因此，可以得出结论：平衡数据能够显著提升性能。我们还测试了使用多模态的效果：在单任务网络中将深度作为额外输入时，准确率提升至24.02%（MMST）。如果在多任务网络中使用多模态输入，平均交并比准确率则进一步提升至24.45%（MMMT-D）。尽管均方根误差表现良好，但需要注意的是，深度作为输入提供，因此该数值只能确保数据不会过拟合。在机器人运行期间，我们假设无法获取深度信息。为符合机器人应用场景，我们将MMMT-D用作教师网络，并将其输出用作伪标签。

我们使用 DNN 最后一层 Softmax 层中取值大于阈值的标签 $\tau = 0.9$。我们使用这份平衡数据训练学生网络，并为缺失数据使用额外的伪标签（UMMT-PS）。平均交并比提升至 25.38%，均方根误差降至 0.6534。这些结果表明，我们通过多任务学习深度神经网络生成伪标签的方法提升了 NYUDv2 数据集上的性能。

这与条件1相符：深度估计和f_1 f_2（语义分割）的子函数具有更多共同子函数（特征），而非不共同的子函数；或者与条件2的镜像情形相符，即相对于f_1 f_3，具有更多共同子函数（特征）。为了解语义f_2分割与深度估计之间的关系属于哪种条件，我们进行了接下来的评估。



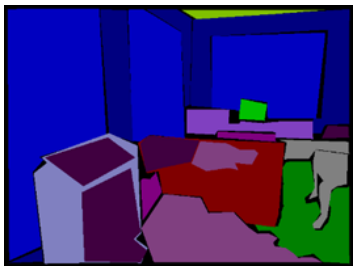
a) NYUDv2示例RGB图像



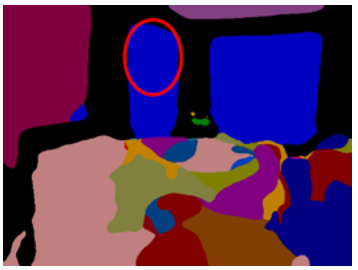
b) 深度的真实值



c) 深度估计 with UMMT-PS



d) GT for 语义分割



e) UMMT-PS 的估计



f) UMMT 的估计

图 5. NYUDv2 数据集中的一个样本图像的定性结果。详情见正文。

我们测试能否采用类似的设置来获取深度伪标签。这里，我们假设有 250 个深度标签和 500 个语义分割标签。由于多模态多任务（MMMT）设置的表现最佳，我们选择该设置来估计伪深度标签。我们使用语义标签和 RGB 作为输入，获得缺失的深度伪标签（MMMT-S）。我们使用这些伪标签训练了一个新网络（UMMT-PD）。然而，结果表明，性能并未提升。这说明深度伪标签的估计并不准确，也就是说，语义分割并不能提升深度分割。结果表明，语义分割与深度估计之间的关系属于条件2，即深度估计的共有子函数（特征）能够使语义分割取得较高的估计性能。但语义分割中与深度估计不共有的子函数（或公式 2 中的 f_{3k} f_{2p} ）更多（）。

$$\|f_{3k}\| > \|f_{3i}\|$$

因此，我们可以得出结论：在这种情况下，将语义分割标签用作输入并不能改善深度估计；但深度能够改善语义分割。

图 *5* 展示了一个定性结果样本。第一行展示了 RGB 图像及其对应的深度真值（GT）和深度估计结果。GT 深度存在噪声，降低了深度估计的准确率。下方展示了语义分割真值（GT）和分割结果。在 UMMT-PS 中，圆圈标出的区域属于单类别区域，而 UMMT 错误地将该区域估计为多类别分割。这个样本展示了使用多任务网络和伪标签带来改进的典型案例。

Cityscapes

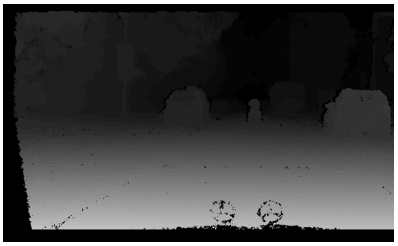
接下来，我们使用 Cityscapes（CS）数据集，在一个更大的自动驾驶车辆采集数据集上验证我们的论点。该数据集包含 2975 张带有深度和语义分割信息的训练图像，以及 500 张验证图像（Cordts 等，2016）。语义标签共有 19 个类别。我们采用两种不同的配置对该数据集进行评估：首先，使用较大的训练集，包含 2975 张图像，并使用 500 张测试图像；其次，使用规模小得多的图像集，即 300 张训练图像和 200 张测试图像，以测试样本数量的影响。

我们采用了与 NYUDv2 实验类似、能够提升性能的策略，并使用了相同的深度神经网络架构。学习率设为 0.0001，批量大小设为 2。我们使用 UMST-S 作为基线。随后，我们采用了多任务设置 UMMT，该设置在前一项实验中提升了平均交并比。然而，在这种情况下，加入深度估计任务反而降低了分割性能，这可能是由于深度标签存在噪声。为了测试伪标签，我们没有使用 UMST-S 中使用的 250 张已标注图像；因此，平均交并比的下降可能归因于已标注图像数量减少。请注意，与 NYUDv2 相比，深度估计指标 **均方根误差** 和 **对数均方根误差** 更高（表现更差）。由于 NYUDv2 是室内数据集，距离较近；而对于 CS，数据采集自室外，距离范围从几米到两百米，因此深度估计的误差率较高。在 MMMT-D 中，我们将深度信息作为网络的输入，该网络将作为最终单模态网络的教师网络。在这种情况下，平均交并比如预期般提升至 73.6 %；然而，深度估计并未显著改善。这可能是由于语义分割的学习率较高。最后，我们使用 MMMT-D 估计得到的额外 250 个伪语义标签，并在 UMMT-PS 的损失计算中将其作为真实标签。正如表5所示，仅使用 RGB 进行语义分割估计和深度估计，性能有所提升。UMMT-PS 的平均交并比在单模态估计中取得了第二好的结果。结果并未像在 NYUDv2 数据集中那样显著提升，这可能归因于样本数量较多。

由于样本数量较多，网络能够学习其参数，并已展现出较高的性能；对于这种深度神经网络架构而言，性能可能只有小幅提升空间。为检验这是否是原因，我们使用有限数量的带标签图像进行实验。对于单模态语义分割任务，我们从训练集中使用300个语义分割标签，并在验证集上进行评估，如UMST-S-L所示（L表示样本数量**较少**）。其性能不及使用大量样本训练的模型，这符合预期。当我们使用多任务模型进行深度估计和语义分割，并以200个语义分割标签和200个深度真实值进行训练（UMMT-L）时，由于语义分割标签数量较少，平均交并比变化不大。这表明多任务学习能够提升网络性能，而且尽管用于语义分割的训练样本较少，性能也不会下降。一个有趣的结果是，与使用较多样本训练的模型相比，深度估计结果有所提升。这可能是因为较小数据集中深度样本的分布不同。当我们将深度作为额外输入提供给模型（MMMT-L）时，语义分割性能有所提升，但深度估计性能没有改善；这与之前使用大量样本时的情况相似。因此，这同样可能归因于语义分割与深度估计的学习率不同。最后，我们再次使用MMMT-L作为教师网络，并将其输出用作语义分割的伪标签，以此训练一个名为UMMMT-PS-L的新模型，该模型仅以RGB作为输入。在图像数量较少时，语义分割性能显著提升。最后，这些发现与先前设置中得到的条件2一致。



a) NYUDv2 RGB图像示例



b) 深度真实值



c) 深度 估计 with UMMT-PS



d) GT for 语义分割



e) UMMT-PS 的估计



f) UMMT 的估计

图6。 Cityscapes数据集中的一个样本图像的定性结果。详情见 tex t.

表5. Cityscapes结果。

	mIoU	rmse	对数均方根误差	输入	Out	loss	||A||	||B||	||v||	||f1||	||f2||
UMMT	71.70	32.65	**10.45**	RGB	语义分割 + 深度	GTS + GTD	2975	-	500	D	S
MMMT-D	**73.60**	*33.80*	*10.40*	RGB + D	语义分割 + 深度	GTS + GTD	3475	2975	500	D	S
UMST-S	72.87	-	-	RGB	S	GTS	2975	-	500	S	S
UMMT-PS	73.10	**32.60**	10.50	RGB	语义分割 + 深度	PSS + GTS + GTD	3475	2975	500	D	S
UMMT-L	52.64	27.92	9.56	RGB	语义分割 + 深度	GTS + GTD	300	200	200	D	S
UMST-S-L	52.83	-	-	RGB	S	GTS	300	200	200	S	S
MMMT-L	53.84	27.70	9.68	RGB+深度	S+D						S
UMMT-PS-L	54.06	29.11	7.84	RGB	语义分割 + 深度	PSS + GTS + GTD	300	200	200	D	S

定性结果如图6所示。首先需要指出，由视差图得到的深度存在明显噪声。深度估计准确率较低可能与此有关。语义分割方面，圈出的区域显示了相关改进。例如，UMMT会在道路中央臆测出一条侧道，而UMMT-PS不会。UMMT会将公交车的一部分识别为汽车，而UMMT-PS不会。最后，UMMT会将汽车后方的一辆卡车识别为汽车，而UMMT-PS则能正确地将其识别为卡车。

结论

我们提出了一种方法，可在数据不平衡场景下为多任务深度神经网络高效利用所提供的全部标签。当所提供的样本数量较少时，使用全部数据尤为重要。我们提出了一种为单模态多任务网络生成伪标签的方法。观察表明，对于包含深度估计与语义分割的双任务，深度估计可通过多模态教师网络提升语义分割任务；然而，这一结论对语义分割任务并不成立。作为

未来工作的一部分，可以对两个以上的任务进行分析，并开展与若干任务相关的实证评估。

作者贡献

第一作者开发了这些方法并进行了实验。

Etik

本文的发表不存在任何伦理问题。

利益冲突

作者声明不存在任何利益冲突。

致谢： 本研究由土耳其科学技术研究理事会通过欧盟委员会地平线2020玛丽·斯克沃多夫斯卡-居里行动联合资助计划的Circulation2计划提供支持。

ORCID

第一作者姓名 https://orcid.org/0000-0002-2436-3186

参考文献

Boursier, E., Konobeev, M., & Flammarion, N.（2022年）。稀缺数据下多任务学习的迹范数正则化。学习理论年度会议（*COLT*），*178*，第1–25页。取自 http://arxiv.org/abs/2202.06742

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.（2002）。SMOTE：合成少数类过采样技术。《人工智能研究杂志》，*30*，321–357。

Choi, H. S., Jung, D., Kim, S., & Yoon, S.（2022年）。通过分类器与生成器之间的协同交互进行不平衡数据分类。《*IEEE*神经网络与学习系统汇刊》，*33*，3343–3356。

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., . . . Schiele, B.（2016）。用于城市场景语义理解的Cityscapes数据集。计算机视觉与模式识别会议。

Crawshaw, M.（2020年）。深度神经网络多任务学习：综述。*arXiv*。检索自 http://arxiv.org/abs/2009.09796

Ghiasi, G., Zoph, B., Cubuk, E. D., Le, Q. V., & Lin, T.-Y.（2021年）。用于学习通用表征的多任务自训练。*ICCV*，（第8856–8865页）。检索自 http://arxiv.org/abs/2108.11353

He, M., Yang, Z., Zhang, G., Long, Y., & Song, H.（2024年）。IIMT-net：利用交互信息进行语义分割和深度估计的 Poly-1权重平衡多任务网络。《图像与视觉计算》，*148*。

Kendall, A., Gal, Y., & Cipolla, R.（2018年）。利用不确定性为场景几何和语义的损失加权的多任务学习。计算机视觉与模式识别会议，（第 7482–7491 页）。

Kokkinos, I.（2017年）。UberNet：使用多样化数据集和有限内存训练通用卷积神经网络，以处理低层、中层和高层视觉。计算机视觉与模式识别会议，（第 5454–5463 页）。

Leang, I., Sistu, G., Burger, F., Bursuc, A., & Yogamani, S.（2020年）。动态任务加权方法 for 自动驾驶系统中的多任务网络。智能国际会议交通系统。Lee, S., Rameau, F., Im, S., & So, I. (2022年). 动态场景中的自监督单目深度与运动学习：语义先验助力。国际计算机视觉期刊, *130*, 2265–2285. Milli, E., Erkent, Ö., & Yılmaz, A. E. (2023). 多模态多任务（3mt）道路分割。*IEEE*机器人与自动化快报。Nathan Silberman, P. K., & Fergus, R. (2012). 基于RGBD图像。*ECCV*。Rottmann, P., Posewsky, T., Milioto, A., Stachniss, C., & Behley, J. (2021年). 改进单目深度通过语义预训练进行估计。*IROS.*Sener, O., & Koltun, V. (2018年). 将多任务学习视为多目标优化。*NeurIPS*,（第527–538页）。Standley, T., Zamir, A. R., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020年). 多任务学习中哪些任务应一起学习？*ICML.* 检索自 http://arxiv.org/abs/1905.07553 Sun, G., Probst, T., Paudel, D. P., Popovic, N., Kanakis, M., Patel, J., . . . Van Gool, L. (2021年). 面向多任务学习的任务切换网络。*ICCV*,（第8291–8300页）。Vandenhende, S., Georgoulis, S., & Van Gool, L. (2020年). MTI-NET：面向多任务学习的多尺度任务交互网络。*ECCV.* 检索自 http://arxiv.org/abs/2001.06902 Vandenhende, S., Georgoulis, S., Gansbeke, W. V., Proesmans, M., Dai, D., & Gool, L. V. (2022年). 面向密集预测任务的多任务学习：综述。*PAMI, 8828*, 1–20. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., . . . 等 (2020年). 面向视觉识别的深度高分辨率表征学习。*IEEE*模式分析与机器智能汇刊, *43*, 3349–3364. Wang, Y., Tsai, Y.-h., Hung, W.-c., Ding, W., Liu, S., & Yang, M.-h. (2022年). 面向语义与深度的半监督多任务学习。*WACV*,（第2505–2514页）。Xu, D., Ouyang, W., Wang, X., & Sebe, N. (2018年). PAD-Net：用于同时进行深度估计与场景解析的多任务引导预测与蒸馏网络。计算机视觉与模式识别会议,（第675–684页）。Xu, Y., Yang, Y., & Zhang, L. (2023). 结合知识蒸馏的密集预测多任务学习。*ICCV.*Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2018年). Taskonomy：解耦task迁移学习。计算机视觉与模式识别会议，（第 6241–6245 页）。