计算机科学
COMPUTER SCIENCE

联合语义分割和深度估计的多任务学习研究

罗会兰, 叶桔

引用本文

罗会兰, 叶桔. 联合语义分割和深度估计的多任务学习研究[J]. 计算机科学, 2023, 50(6A): 220100111-10.

LUO Huilan, YE Ju. Study of Multi-task Learning with Joint Semantic Segmentation and Depth Estimation [J]. Computer Science, 2023, 50(6A): 220100111-10.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

联合人体姿态估计和多目标跟踪的跨数据集学习

Cross-dataset Learning Combining Multi-object Tracking and Human Pose Estimation

计算机科学, 2023, 50(6A): 220400199-7. https://doi.org/10.11896/jsjkx.220400199

注意力特征融合的孪生网络目标跟踪方法

Attentional Feature Fusion Approach for Siamese Network Based Object Tracking

计算机科学, 2023, 50(6A): 220300237-9. https://doi.org/10.11896/jsjkx.220300237

基于边缘优化和全局建模的多路径语义分割

Multi-path Semantic Segmentation Based on Edge Optimization and Global Modeling

计算机科学, 2023, 50(6A): 220700137-7. https://doi.org/10.11896/jsjkx.220700137

基于CT图像语义的COVID-19实例分割与分类网络

COVID-19 Instance Segmentation and Classification Network Based on CT Image Semantics

计算机科学, 2023, 50(6A): 220600142-9. https://doi.org/10.11896/jsjkx.220600142

基于数据融合的半监督高分遥感影像语义分割

Semi-supervised Semantic Segmentation for High-resolution Remote Sensing Images Based on DataFusion

计算机科学, 2023, 50(6A): 220500001-6. https://doi.org/10.11896/jsjkx.220500001

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100111

联合语义分割和深度估计的多任务学习研究

罗会兰　叶　桔
江西理工大学信息工程学院　江西 赣州 341000

摘　要　语义分割和深度估计任务是对图像像素级分类的研究,是两个高度相关的任务。从共享特征学习和特征交互融合两个角度出发,提出两个不同的多任务学习架构,即基于压缩激励模块(Squeeze-and-Excitation,SE)和金字塔池化的多任务学习网络(Multi-task Learning with SE and Pyramid Pooling,MTL_SPP),以及基于压缩激励和可选择权重(Selective Weight,SW)的多任务学习网络(Multi-task Learning with SE and Selective Weights,MTL_SSW),来联合学习语义分割和深度估计。MTL_SPP架构由共享骨干特征网络和任务特定的子网络组成,利用SE模块构建任务特定子网络,并利用金字塔池化增强特征提取。MTL_SSW在MTL_SPP的基础上,让任务特定子网络的语义分割特征和深度估计特征通过SW模块进行相互指导和优化,学习对特定任务更具判别性的特征。实验结果表明,提出的两种方法在NYUD_v2和SUNRGBD两个数据集上获得了优于先进方法的效果。

关键词:多任务学习;语义分割;深度估计;压缩激励;可选择权重
中图法分类号　TP391

Study of Multi-task Learning with Joint Semantic Segmentation and Depth Estimation

LUO Huilan and YE Ju
School of Information Engineering,Jiangxi University of Science and Technology,Ganzhou,Jiangxi 341000,China

Abstract　Semantic segmentation and depth estimation are two highly related tasks of image pixel-level classification. This paper proposes two different multi-task learning architectures from the perspectives of both shared feature extraction and feature interaction fusion:multi-task learning with SE and pyramid pooling (MTL_SPP) based on the squeeze and excitation (SE) and pyramid pooling,and multi-task learning network (MTL_SSW) based on se and selective weights (SW) to jointly learn semantic segmentation and depth estimation. The MTL_SPP architecture consists of shared backbone feature network and task-specific sub-networks,using the SE module to construct task-specific sub-networks and pyramid pooling to enhance feature extraction. Based on MTL_SPP,MTL_SSW adds SW modules which allows the semantic segmentation features and depth estimation features from task-specific sub-networks to guide and optimize each other, o it can learn more discriminative features. Experimental results show that the two proposed methods obtain better results than the state-of-the-art methods on both NYUD_v2 and SUNRGBD datasets.

Keywords　Multi-task learning,Semantic segmentation,Depth estimation,Squeeze and excitation,Selective weights

1　引言

多任务学习[1](Multi Task Learning,MTL)同时学习多个相互有关联的预测任务,旨在共享相关信息,提高各个任务的泛化性。MTL利用多个学习任务中所包含的有用信息来帮助各个任务学习得到更为准确的学习器,例如文献[2-4]使用多任务学习同时提高语义分割和表面法线检测等任务的性能。此外,有些多任务学习研究构造辅助任务来提高主任务性能,例如文献[5]利用语义分割任务作为辅助信息来提高深度估计的准确度,文献[6-7]利用边缘检测和表面法线检测作为辅助任务来提高语义分割和深度估计任务的性能。多任务学习在现实世界的场景理解中有比较多的应用,例如:自动驾驶[8-9]需要对场景中的所有对象进行定位,并估计他们的距离和轨迹;虚拟现实[10]需要为用户提供实时的互动效果,同时进行定位和运动轨迹的估计。

语义分割和深度估计是场景理解的两个重要视觉任务,都需要对每个像素进行预测。不同的是,语义分割是给出每个像素的类别标签,得到图像的语义信息。深度估计是估计图像中每个像素相对于拍摄源的距离,从而描述物体的空间几何位置关系。在图像的场景理解中,语义分割信息和物体间的几何位置信息有很强的关联性,所以联合训练语义分割和深度估计两个任务可以提升各自的估计准确性。典型的

基金项目:国家自然科学基金(61862031,61462035);省级学位与研究生教育教学改革研究项目重点项目(JXYJG-2020-120);江西省教育厅科学技术研究项目(GJJ200859,GJJ200884)
This work was supported by the National Natural Science Foundation of China(61862031,61462035),Education Teaching Reform Research Project of Jiangxi Province(JXYJG-2020-120) and Science and Technology Research Project of Jiangxi Provincial Education Department(GJJ200859,GJJ200884).
通信作者:罗会兰(825667187@qq.com)

联合学习此两类任务的方法是，通过学习共享特征，利用任务间的潜在关系，达到特征信息的相互补充和相互约束的目的，从而提高各个任务的性能。如文献[5]提出一个统一的联合预测框架，引入分层条件随机场，将语义分割和深度估计结合起来，实现联合优化。文献[11]在共享编码器网络基础上，使用3个并行解码器分别学习语义分割、实例分割和深度估计3个任务。Liu等[2]提出多任务注意网络(Multi-Task Attention Network，MTAN)用于语义分割、深度估计和表面法线预测任务，3个任务共享一个浅层网络，接着用3个自注意力模块分别学习各任务的具体特征，屏蔽对各自任务没有用的信息，提高了语义分割和深度估计等任务的性能。

另一类结合语义分割和深度估计的多任务学习方法通过共享预测结果来提高任务特征学习的效果。PADnet[6]使用多模态蒸馏整合深度估计、边缘检测和表面法线等任务的中间预测结果，利用得到的多模态信息来获得性能的提升。Simon等[7]在PADnet[6]的基础上，使用特征传播模块来细化语义分割、深度估计、边缘检测和表面法线检测4个初始预测结果，实现任务之间信息的交互，最后使用多模态蒸馏模块得到最终的语义分割和深度估计任务的预测结果。

以上的方法虽然在语义分割和深度估计任务上取得了不错的效果，但仍存在一些不足。首先，共享浅层特征的方法未能有效地挖掘不同模态高层语义之间的相关性。另一方面，共享预测结果的方法使用多模态信息来丰富特征，但是其性能会受到预测结果准确性的影响，有时可能会对任务性能有消极的作用。针对这些问题，本文提出了一个基于压缩激励模块和金字塔池化的多任务学习网络MTL_SPP联合学习语义分割和深度估计任务。利用SE模块构建不同的任务分支，不但学习两个任务在浅层基础特征间的共享关系，还学习它们在深层语义特征间的互补关系，并用金字塔池化(Pyramid Pooling Module，PP)来处理最后一个SE模块的输出，学习任务特征的上下文信息，简单有效地改善了任务预测结果。为了更好地实现任务间的高层信息共享，受SKNet[12]的启发，在MTL_SPP基础上，本文进一步提出了一个基于压缩激励和可选择权重(Selective Weight，SW)的多任务学习模型MTL_SSW联合学习语义分割和深度估计，利用可选择权重模块选择性地融合任务间的部分特征，有效利用任务之间的互补信息，并用空洞空间金字塔池化模块(Atrous Spatial Pyramid Pooling，ASPP)提取更密集的特征信息，从而提高语义分割和深度估计的性能。

本文的主要贡献如下：

(1)提出了基于压缩激励模块和金字塔池化的单流多任务学习框架MTL_SPP。在共享深度骨干网络特征的基础上，利用压缩激励模块建立两个任务特定子网络分支，以端到端方式学习深度估计和语义分割任务，同时利用金字塔池化模块加强语义特征和深度特征的上下文信息，获得了更好的平均性能。

(2)在MTL_SPP的基础上，提出了基于可选择权重的多任务学习网络MTL_SSW，在任务特定子网络间使用可选择权重模块进行分层式的特征融合，协同优化语义特征和深度特征，用空洞金字塔池化模块获取更密集的任务特征信息，以较低的计算复杂度获得更高的精度，有效实现了跨任务特征交互，进一步提高了语义分割和深度估计的性能。

2 相关工作

2.1 深度估计

近年来，基于深度学习的单任务模型方法[13-17]有效提升了深度估计的准确率。传统的基于深度卷积网络的深度估计方法通过设计不同的多尺度网络来获得丰富的深度信息，从而对深度特征进行细节优化。文献[13]设计了一个多尺度网络框架，通过3个级联的子网络逐步细化深度特征。Eigen等[14]设计了两个尺度的级联子网络对深度进行回归，前一个子网络得到粗糙的全局场景估计，与原始图片一起输入到另一个子网络中进行优化。进一步地，文献[17]提出多尺度网络结构，结合空间离散策略，使深度特征离散化，在加快模型收敛速度的基础上提高了深度估计的预测准确度。Liu等[18]提出多尺度注意力导向网络，引入空间注意力来提升不同物体类间对比度以增强图像的边界细节。由于从单幅图像中提取的几何信息是有限的，所以最近有些研究联合语义信息和表面法线信息，学习它们与深度信息的相关性，在基于编解码结构的基础上，协同优化深度估计的预测结果。Yin等[16]在共享编解码网络基础上设计了一个简单的损失项，结合虚拟法线的几何信息来提高深度估计预测。文献[15]在共享编码器基础上，用两个解码器分别处理深度特征和语义特征，利用自监督语义信息指导深度估计来处理动态类对象，其语义分割产生的掩码使运动的物体不参与光度损失的计算，从而提升了深度估计的性能。

2.2 语义分割

近年来，基于卷积神经网络(Convolutional Neural Network，CNN)的图像语义分割方法[19-22]取得了很大的进展。Long等[19]提出不含全连接层的全卷积网络(Fully Convolutional Networks，FCN)来适应任意大小的输入，并采用扩张卷积运算增大感受野，从而输出更精细的分割结果。DeepLab[20]将深度神经网络的最后一层响应输入到密集条件随机场(Conditional Random Field，CRF)捕获像素值之间的长范围依赖关系，改善了分割准确度。Unet[21]和Deeplabv3+[22]采用编解码结构，融合底层和高层的信息来预测分割掩码。为弥补高层低分辨率语义特征缺乏图像细节信息和浅层高分辨率特征缺乏语义信息的不足，有些方法[23-26]尝试通过整合上下文信息来提高特征的表达能力。文献[25]中提出ASPP模块，通过使用不同扩张率的扩张卷积来捕获不同感受野的上下文信息。PSPNet[26]引入金字塔池化，建立不同大小的子区域金字塔，并将对其进行池化操作后的结果作为该子区域的上下文信息。ParseNet[23]利用全局平均池化来获取全局的上下文信息。此外，文献[27]在BiseNet[28]的基础上增加全局路径分支，以增强不同尺度的上下文语义信息。Zhao等[29]提出逐点空间注意力网络，使用预测的注意力特征图来指导上下文信息的收集。进一步，文献[30]构建了上下文注意力卷积层来学习点云的细粒度局部特征，利用上下文循环神经网络编码机制来捕捉多尺度局部区域之间的上下文特征。随着深度传感器的广泛应用，一些研究结合RGB数据和深度数据训练模型，有效提高了语义分割的精度。Duan等[31]设计了一个跨模态注意力机制，让RGB特征和深度特征学习互补信息。Seichter等[32]提出一种RGB-D分割方法，同时将RGB和深度图像作为输入，利用深度图像提供的几何

信息来缓解室内场景的复杂性，从而获得更好的分割性能。

2.3 多任务学习

当前的多任务学习方法主要是通过设计特定的网络结构来实现任务间的信息共享，达到提高各任务性能的目的，按网络结构可分为单流多任务学习[2,7,33-35]和多流多任务学习[3-4,36]。单流多任务学习在共享特征提取网络的基础上，为不同的任务构建不同的分支，来达到训练多个任务的目的。MaskRcnn[37]在 Faster-RCNN[38]的基础上增加了一个掩码分支，实现了语义分割和目标检测任务。在利用单流多任务学习进行语义分割和深度估计联合学习的研究中，Kendall 等[11]在一个共享编码网络的基础上，为各任务构造解码分支，利用最大化高斯似然法来调节多任务损失函数中各任务的权重，在语义分割、实例分割和深度估计的多任务场景理解中得到了较好的效果。Liu 等[2]提出一种基于共享浅层特征的多任务注意网络(Multi-Task Attention Network，MTAN)，用 3 个自注意力模块分别学习各任务的具体特征，并提出动态权重平均(Dynamic Weight Average，DWA)策略，通过考虑每个任务的损失变化，在训练过程中自适应任务权重，有效提高了语义分割、深度估计和表面法线检测任务的性能。不同于上述学习共享特征的方法，PADnet[6]则是在共享浅层网络的最后一层产生深度估计、表面法线检测、语义分割和边缘检测 4 个任务的初始预测，然后通过多任务蒸馏来提高深度估计和语义分割任务的最后性能。MTI_Net[7]在 PADnet 的基础上，利用多尺度传播和空间注意力来学习语义、表面法线、边缘信息和深度任务之间的高层语义关系，从而提高语义分割和深度估计的性能。

多流多任务学习是通过在不同任务的网络模型间设计共享结构来选择部分信息进行共享，达到了任务间既相对独立又相互共享的目的。经典的交叉缝合方法(cross-stitch)[36]为不同任务设计了独立的、结构相同的主干网络，然后在不同的网络层之间通过线性组合将不同任务的部分信息进行融合学习，再反馈到各任务网络中进行共享，提高了预测准确度。Gao 等[3]在 cross-stitch[36]的基础上提出判别降维(Neural Discriminative Dimensionality Reduction，NDDR)的分层特征融合方案，自动融合语义分割和表面法线检测任务的不同层特征。2020 年，他们又在 NDDR_CNN 的基础上引入神经结构搜索(Neural Architecture Search，NAS)[4]，为特定任务搜索最佳共享连接结构，在语义分割和表面法线多任务学习的实验中实现了精度提升。

单流多任务学习网络虽然学习任务间潜在关系的能力较低，但网络结构相对简单，容易扩展到任务数量较多的场景中，网络性能受损失权重分配不平衡的影响也会随着任务数量的增加有所降低[2]。而多流多任务网络学习能力相对较强，但网络参数量会随任务数的增加而成倍增加[3-4]。本文综合了单流和双流结构的特点，在共享多级骨干特征提取网络基础上，为语义分割和深度估计设计特定的特征选择分支，同时在任务特定层之间设计交互模块来学习任务间的相关信息，发掘任务间隐藏的有效信息来提高最终的预测效果。

3 提出的方法

为了更好地利用语义分割和深度估计任务间的高层语义关系，本文提出了两种多任务学习框架来联合训练语义分割和深度估计任务：基于 SE 和金字塔特征池化的多任务学习网络 MTL_SPP，以及基于可选择权重的多任务学习网络 MTL_SSW。

3.1 MTL_SPP

MTL_SPP 首先通过一个共享特征提取骨干网络，比如残差网络 resnet50[39]，来学习多级共享特征池；在此共享特征池基础之上，将各级特征输入任务特定的 SE 模块，通过注意力机制学习任务特定的特征。与此同时，选择学习后的浅层特征与更高一级的共享特征共同输入下一个任务特定的 SE 模块，以此来学习各级特征的任务特定的融合方式。在学习到的融合特征基础上，设计了一个金字塔池化模块用于增强特征的上下文信息，最后上采样得到最终的语义分割和深度估计结果。MTL_SPP 的体系结构如图 1 所示，共享骨干网络设计了 4 个残差块，每个残差块输出的特征通过跳跃连接直接输入到任务特定网络。任务特定网络利用 4 个级联的 SE 模块连接对应共享骨干网络的 4 个残差块。最后用金字塔池化模块连接最后一个 SE 注意模块，并进行一个上采样操作，利用双线性插值恢复图像的分辨率。

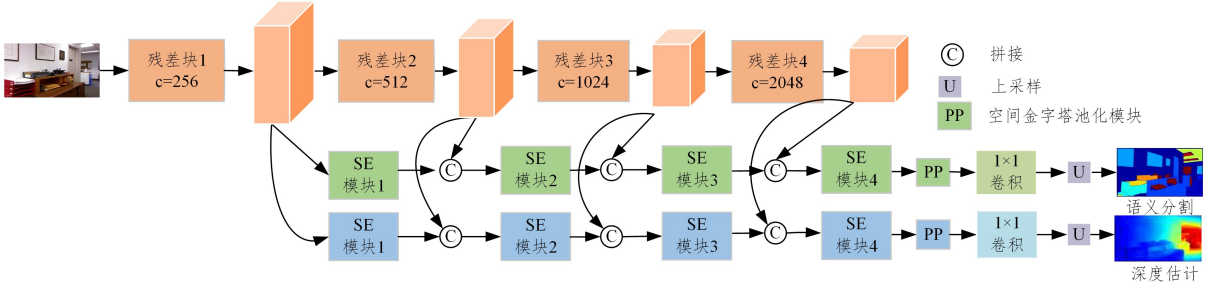


图 1 MTL_SPP 结构图

Fig. 1 Diagram of MTL_SPP structure

在 MTL_SPP 网络中，共享网络为语义分割和深度估计任务学习一个包含浅层细节特征和深层语义特征的共享多级特征池，而任务特定网络可以看成是共享网络的特征选择器，以端到端的方式自动学习特定任务的信息。MTL_SPP 方法的动机是，通过共享多级特征池，学习语义分割和深度估计任务间的潜在语义关系，为彼此提供信息的补充和约束，从而提高两个任务的性能。另一方面，通过为不同任务设计多级特征选择融合学习网络，强化任务特定的信息和学习任务特定的多级特征融合方式，从而避免了任务目标冲突引起的性能下降问题。接下来将分别对 MTL_SPP 中的 SE 模块和金字塔池化模块进行详细论述。

3.1.1 SE 模块

语义分割任务倾向关注图像的颜色和纹理区域，深度估计任务则更倾向关注深度值发生变化的区域，两类任务联合训练可以使得共享骨干网络学习到更为丰富的多级特征。此外，语义分割和深度估计是非常相关的两类任务，同一物体的深度值很大可能是相同或连续的，深度值发生急剧变化的地方很有可能是分割边界，所以它们的联合训练也可以相互提供指导信息和互补信息。为了学习共享特征池中语义分割和深度估计的补充信息来进行任务的指导和优化，同时让任务特定网络以多级融合的方式给予自身任务更多关注，本文设计了由 4 个级联 SE 模块组成的任务特定网络，通过每个 SE 模块学习一个激励特征响应，以注意力的方式聚焦式学习特定任务的多级融合特征。

4 个任务特定的 SE 模块的连接方式如图 1 所示。任务特定的网络的第一个 SE 模块将主干网络的第一个共享残差块的输出特征作为输入。后续的 SE 模块则是将上一级任务特定 SE 模块的输出特征与更高一级的共享特征串联得到的特征作为输入。SE 模块通过生成通道维度的激励响应，以注意力的方式来学习利用共享特征池中的特定信息，和任务特定的特征融合方式，从而学习到任务特定激励特征，提高各任务性能。语义分割任务分支和深度估计任务分支通过 SE 模块，不仅可以学习任务特定通道中的相关信息，使自身任务受到更多的关注；更重要的是，它们的联合训练也迫使多级共享特征池学习任务间相关的细节信息和语义信息，从而 SE 模块可以通过注意力激励响应，利用这些信息进行特定任务的指导和优化。

压缩激励模块 SE 的结构如图 2 所示，SE 模块分压缩和激励两个操作。SE 模块学习共享特征池和任务特定分支的融合特征，并通过压缩激励的方式产生一个注意力值，来激励任务学习对自身更有利的信息。SE 模块首先对共享特征池和上一级任务特定 SE 模块的输出特征进行串联，接着使用 1×1 卷积在对通道降维的同时，学习不同通道特征间的融合交互信息，最后输入到 SE 模块中。这个过程如式(1)所示。

$$U=f^{1\times1}(Concat(X^{i}_{\text{share}},X^{i-1}_{\text{SE}})),i=2,3,4 \tag{1}$$

其中，$f^{1\times1}$ 为标准的卷积操作，卷积核大小为 1×1；X^{i}_{share} 表示第 i 个共享特征池的输出特征图，X^{i-1}_{SE} 表示第 $i-1$ 个 SE 模块的输出特征图。

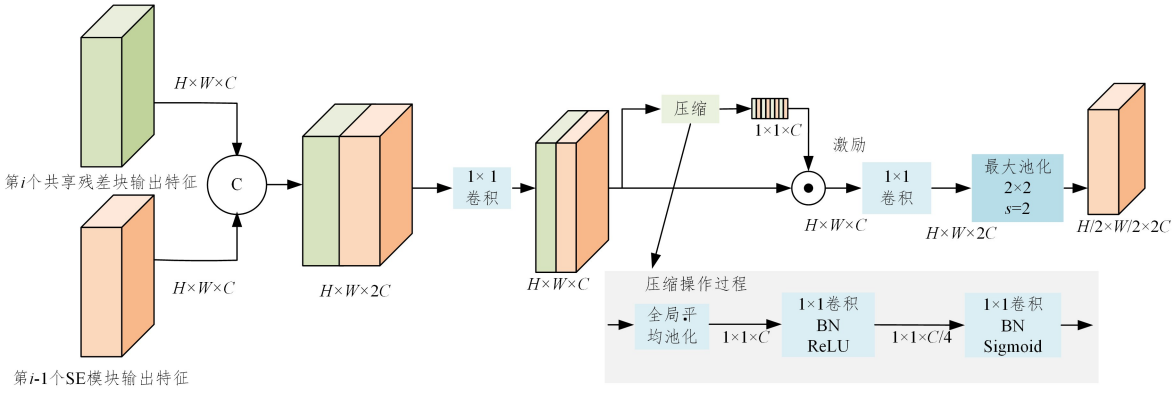


图 2 SE 模块结构图

Fig. 2 Diagram of SE module structure

SE 模块首先在空间维度上对输入特征 U 进行压缩，利用空间全局平均池化得到初始通道激励值 Z_c。特征图 U 中包含了从自身任务学习的特征和共享特征池的特征，所以经过压缩操作得到的通道激励值 Z_c，不仅反映了通道上的共享特征全局分布，而且具有任务特定的全局感受野。初始通道压缩激励值 Z_c 的计算过程如式(2)所示：

$$z_C=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}U_C(i,j) \tag{2}$$

其中，$U_C(i,j)$ 表示任意通道 c 的特征图，H 和 W 表示特征图的大小。

为了捕获特定任务通道方面的依赖关系，在得到初始通道压缩激励值 Z_c 后，设计了两个连续的 1×1 卷积，进一步学习通道间非线性的交互关系。其中，第一个 1×1 卷积层将通道数进行压缩，经过 ReLU 层的非线性化处理后，输入到第二个 1×1 卷积层，恢复原来的通道数。最后使用 Sigmoid 激活函数，得到增强的通道激励响应 Z_c，过程如式(3)所示。

$$Z_c=\text{Sigmoid}(BN(f^{1\times1}(\text{ReLU}(BN(f^{1\times1}(z_c)))))) \tag{3}$$

其中，$f^{1\times1}$ 为 1×1 卷积，BN 为批正则化操作，ReLU 为线性整流操作，Sigmoid 为激活函数。

激励操作则是将增强后的激励响应和特征图 U 进行逐通道相乘操作，得到任务特定激励特征 $S\in R^{H\times W\times C}$，学习到更具区分度和抗噪性能的特征。这个过程如式(4)所示：

$$S_c=Z_c\odot U_c \tag{4}$$

其中，U 表示特征图，$\odot$ 表示对应位置元素相乘。然后将 SE 的输出特征进行 1×1 卷积操作，在通道上进行升维，实现特征的进一步融合学习。接着使用步长为 2 的 2×2 最大池化操作，进行下采样，得到 SE 模块的最后输出。所以 SE 模块的最后输出与更高一级的共享残差块的输出特征图具有相同的分辨率大小。这种先在通道上升维，后在空间上压缩的操作，可以在引入极少量参数的情况下，进一步学习空间位置信息。这个过程如式(5)所示：

$$S=f^{1\times1}(Down(S)) \tag{5}$$

其中，$f^{1\times1}$ 为 1×1 卷积操作，Down 表示下采样，这里使用了步长为 2 的 2×2 的最大池化操作。

3.1.2 金字塔池化

对于语义分割和深度估计任务来说，不同物体需要在不同的感受野下才能获得完整的特征信息，从而实现准确的分割和估计。所以在 4 个级联 SE 模块学习到任务注意力融合特征后，本文采用金字塔池化模块进一步学习不同感受野的

上下文信息，对 SE 模块的输出特征进行增强处理，结构如图 3 所示。

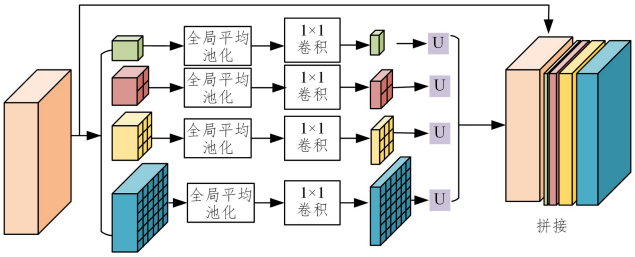


图 3　金字塔池化模块

Fig. 3　Pyramid pooling module

首先通过不同尺度的全局平均池化得到 4 种尺度不同的子区域特征信息。实现过程如式(6)所示，即首先将图像划分成不同数量的子区域，然后在每个子区域中应用全局平均池化。

$$X_p = Ada^p(S), p = 1,2,3,6 \tag{6}$$

其中，Ada^p 表示对不同数量的子区域进行全局平均池化；p 表示将原特征图划分成不同数量的子区域，即分别为 1×2，2×2，3×3 和 6×6 个子区域；S 表示最后一个任务特定的 SE 模块的输出特征。

为了在任务特定层中利用这些不同区域、不同尺度的全局上下文信息，在得到不同尺度特征后，分别使用了 1×1 的卷积对它们进行通道数降维，然后通过双线性插值的上采样操作将它们恢复到原输入特征大小。接着使用拼接操作将 4 种包含不同尺度感受野的特征与原输入特征 S_T 进行拼接，得到最后的输出 X_{out}，这个过程如式(7)所示。全局金字塔池化通过学习不同尺度子区域的全局信息及它们间的关系，来增强特定任务的上下文信息，从而提高语义分割和深度估计两个任务的性能。

$$\begin{cases} X_{out} = \mathrm{Concat}(S_T, X_{(out,1)}, X_{(out,2)}, X_{(out,3)}, X_{(out,6)}) \\ X_{out,p} = UP(f^{1\times1}(X_p)) \end{cases} \tag{7}$$

其中，$f^{1\times1}$ 表示核大小为 1×1 的卷积操作，$X_{(out,p)}$ 表示尺度为 $p\times p$ 的全局平均池化分支的输出特征。

3.2　基于可选择权重的多任务框架 MTL_SSW

本文提出的联合语义分割和深度估计的多任务学习框架 MTL_SPP，虽然可以利用 SE 模块从共享特征池选择学习任务特定特征，并以多级特征融合的方式聚焦式学习特定任务的信息，解决由于任务目标冲突而引起的性能下降问题，但任务特定 SE 模块学习到的特征只能在自身子网络中进行传递，不能很好地利用另一个任务特定网络中学习到的特征信息来补充指导自身任务的学习。所以本文在 MTL_SPP 的基础上，提出一个基于可选择权重的多任务框架 MTL_SSW，旨在利用 SE 模块学习共享特征池多级特征的同时，让任务特定子网络的语义分割特征和深度估计特征通过 SW 模块进行选择性的融合和分离特征，从而实现任务间的相互指导和优化。

MTL_SSW 的网络架构如图 4 所示，它在 MTL_SPP 基础上，在两个任务特定网络间增加了可选择权重模块(Selective Weight，SW)，以加强任务间的交互和学习。不同任务 SE 模块的同一级别的输出特征，首先输入 SW 模块进行融合交互学习，旨在从其他任务中学习互补信息；然后将经过 SW 模块信息交流后的输出特征与更高一级的共享骨干特征串联后，一起输入到下一层的任务特定 SE 模块中学习。MTL_SSW 还采用了空洞空间金字塔池化模块(ASPP)来代替 MTL_SPP 的金字塔池化模块，以获取更密集的特征信息，缓解上采样特征带来的噪声映射。MTL_SPP 中只利用了共享多级特征池来隐式学习语义分割和深度估计任务间潜在的语义关系；MTL_SSW 则进一步利用 SW 模块，以分层选择融合的方式，显式学习任务特定 SE 模块中两个任务间的互补特征信息，强化任务特定网络学习的相互指导和优化。总而言之，MTL_SSW 不仅利用 SE 模块从共享多级特征池中学习到适合特定任务的融合特征，而且通过 SW 模块建模语义分割和深度估计任务子网络之间的依赖关系，利用可选择权重融合语义分割特征和深度估计特征之间的相关通道，并利用 ASPP 模块对密集特征进行学习，帮助任务特定网络学习更具判别性的特征信息。

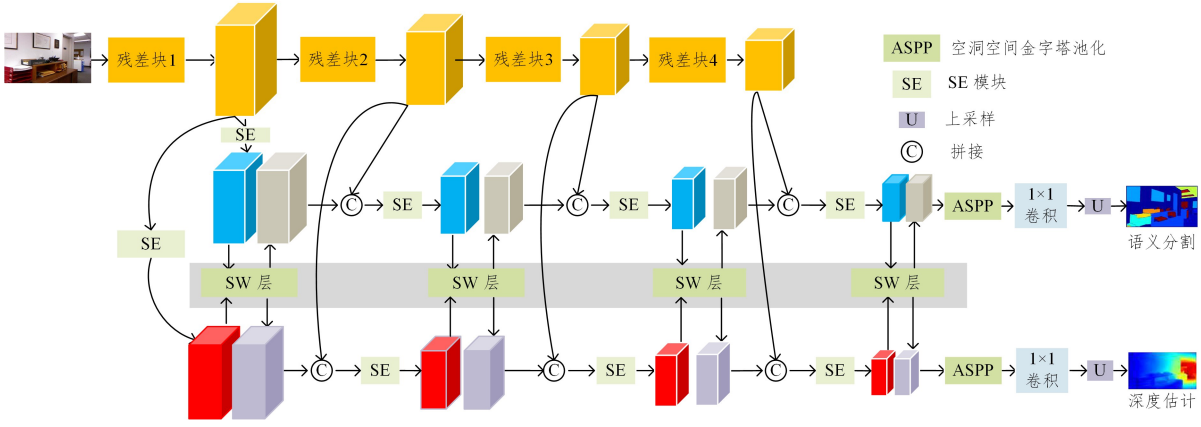


图 4　MTL_SSW 多任务学习模型

Fig. 4　MTL_SSW multi-task learning model

3.2.1　SW 模块

可选择权重 SW 模块的结构如图 5 所示，SW 模块主要通过“拼接融合-选择权重-分离降维”3 个操作完成。其设计目的是让任务特定网络中学习到的深度估计特征和语义分割特征进行融合和交互后，输出特定任务的最佳判别性特征，让语义分割信息和深度估计信息进行相互指导和优化，达到提高任务预测效果的目的。

“拼接融合”操作利用从不同任务的 SE 模块中学习到的层次特征，首先对相同级别的不同任务特征进行拼接，拼接后的特征包含了语义分割任务分支和深度估计任务分支的特征。“选择权重”操作通过为不同任务设计一个可训练权重矩阵来学习不同任务间信息的融合和交互方式，如式(8)所示。

$$Y_{t1}=[X_{t1},X_{t2}]W_1^{\mathrm{T}}$$
$$Y_{t2}=[X_{t1},X_{t2}]W_2^{\mathrm{T}} \tag{8}$$

其中，W 表示可学习的参数矩阵，X_{t1} 和 X_{t2} 分别表示语义分割和深度估计分支的输入特征。

“分离降维”操作利用学习到的任务特定权重矩阵，通过如式(8)的操作，得到任务特定的特征输出，实现特征的分离，然后利用 1×1 的卷积对它们进行通道数降维，并进一步学习通道特征间的关系。因此经过 SW 模块后的语义特征学习到了深度信息，能更有效地对不同物体进行区分，而深度特征能利用语义特征中的纹理信息学习有效定位，实现两个任务的相互学习。

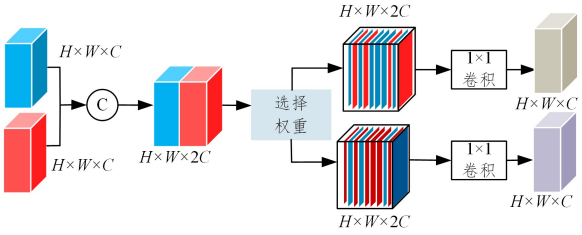


图 5 SW 模块

Fig. 5 SW module

3.2.2 空洞空间金字塔模块

MTL_SSW 框架采用了空洞空间金字塔池化模块(Atrous Spatial Pyramid Pooling，ASPP)来增强各任务特定网络分支最终学习到的特征，它的结构如图 6 所示。

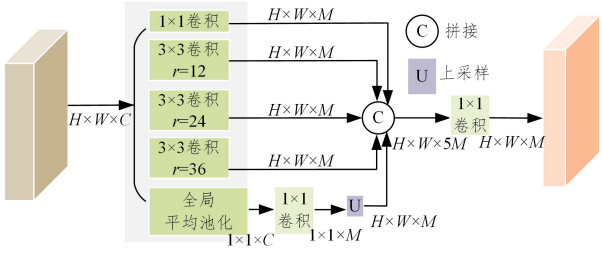


图 6 ASPP 模块

Fig. 6 ASPPmodule

ASPP 通过使用不同扩张率的滤波器来学习不同尺度下的上下文信息。相较于 MTL_SPP 中的金字塔池化模块，ASPP 通过执行不同尺度的全局池化，再上采样来获取多尺度信息，能有效避免上采样操作引入的噪声映射。另一方面，ASPP 通过使用不同扩张率的空洞卷积进行特征的学习，来捕获不同感受野下的区域信息，从而学习每个位置在不同感受野下的上下文特征，它更加适合于多模态信息交互特征的增强。本文的消融实验也证实了，在 MTL_SSW 框架中，使用 ASPP 能获得比金字塔池化更好的特征增强效果。

4 实验

4.1 实验设置

(1)数据集

本文使用常用的 NYUD-V2 和 SUNRGBD 室内数据集来评估所提出的方法。NYUD-V2 数据集含有语义和深度等多个任务标签，可同时用于深度估计和语义分割等任务的训练，语义标签共有 40 类，深度图的缺失值利用 Microsoft Kinect 采集到的数据进行了填充。NYUD-V2 数据集包含 1449 张原始图片，其中 795 张图片用于训练，654 张图片用于测试，图片大小调整为 480×640。SUNRGBD 数据集也具有语义和深度值标签，可用于语义分割和深度估计任务，语义标签共有 37 类，共包含 10335 张图片，其中 5285 张图片用于训练，5050 张图片用于测试。

(2)损失函数和评价指标

由于实验需要同时训练两个任务，本文对语义分割使用 softmax 交叉熵损失函数，对深度估计使用 L1 损失函数，它们的加权和作为网络的总损失，如式(9)所示。

$$L_{\mathrm{tol}}=\sum_{t=1}^{T}W_t * L_t \tag{9}$$

其中，t 表示不同的任务，L_t 表示不同任务的损失，W_t 表示权重。在本文的实验中，语义分割和深度估计的损失权重均设置为 1。

语义分割的性能度量采用了平均交并比(Mean Intersection over Union，MIOU)进行评估，MIOU 值越大表明性能越好。深度估计采用了均方根误差(Root Mean Squared Error，rmse)和对数均方根误差(Mean log10 Error，log_rmse)来验证算法性能，误差值越小表示性能越好。

(3)训练

本文使用在 ImageNet 数据集上预训练过的 resnet50 网络作为本文的主干网络，并用扩张卷积代替残差网络的最后一层下采样，让最后一层的网络分辨率为原始分辨率的1/16，从而保留图像的细节信息。本文使用 ADAM 优化算法训练网络，初始学习率设置为 0.0001，并使用 Poly 策略自适应调整学习率的方法来平衡网络学习速度，batch_size 设为 8。

4.2 与最先进方法比较

为了验证本文提出的方法在联合学习语义分割和深度估计任务上的有效性，将其与近年来先进的多任务算法在 NYUD-v2 数据集和 SUNRGBD 数据集上进行了定量比较。参与比较的先进方法中，cross_stitch[36]，MTAN[2] 和 NDDR_CNN[3] 等多任务框架都是基于共享特征的结构，其中 NDDR_CNN 和 Cross-stitch[36] 为双流多任务模型，MTAN[2] 是一个基于编解码结构的单流多任务模型。多任务框架 PADnet[6] 和 MTI_Net[7] 是基于共享预测结果的结构。为了公平比较，本文与其他先进方法使用了同样的骨干网络 resnet50。

(1)在 NYUD-v2 数据集上的对比结果

图 7 为本文方法与其他先进方法在语义分割和深度估计任务上的性能散点图，其中横坐标表示深度估计的性能，纵坐标表示语义分割的性能，不同的点表示不同的方法，坐标越往右上角性能越好。在图 7 的左边子图中，深度估计采用均方根误差 rmse 作为度量指标，而右边子图深度估计采用的性能度量是对数均方根误差 log_rmse。从图 7 的结果可以看出，本文提出的两种多任务学习框架得到了比其他先进方法更好的综合性能，而且 MTL_SSW 比 MTL_SPP 的平均性能更好。

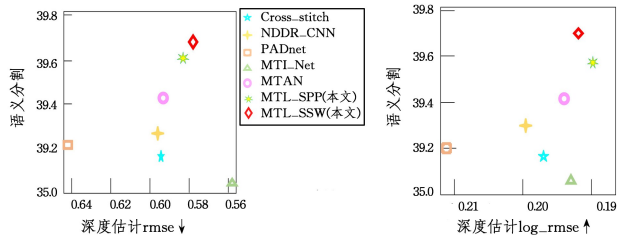


图 7 在 NYUD-v2 数据集上的性能比较

Fig. 7 Performance comparison on NYUD-v2 dataset

表 1 列出了本文方法和其他先进方法在 NYUD-v2 数据集上的语义分割和深度估计定量比较结果。从表 1 的实验结果可以看出，本文提出的 MTL_SPP 和 MTL_SSW 在 NYUD-v2 测试集上的语义分割平均交并比分别为 39.611 9% 和 39.879 6%，MTL_SPP 深度估计的 rmse 和 log_rmse 分别为 0.584 和 0.192 8，MTL_SSW 深度估计的 rmse 和 log_rmse 分别为 0.5755 和 0.1904。相比于双流多任务模型 NDDR_CNN 和 cross_stitch，MTL_SSW 的 MIOU 比 NDDR_CNN 高 0.549 5%，比 cross-stitch 高 1.129 9%；rmse 比 NDDR_CNN 低 0.018 8，比 cross_stich 低 0.021 8。与单流多任务模型 MTAN 相比，MTL_SSW 和 MTL_SPP 对于语义分割任务，MIOU 分别提高了 0.410 7% 和 0.1%，对于深度估计任务，rmse 分别降低了 0.019 8 和 0.015 3。所以，与上述基于共享特征的多任务结构相比，本文的方法在 NYUD-v2 数据集上取得了最优的性能。

表 1　与其他先进方法在 NYUD-v2 数据集上的实验对比

Table 1　Experimental comparison of the proposed method with other advanced methods on NYUD-v2 dataset

方法	语义分割	深度估计		参数
	MIOU ↑ /%	rmse ↓	log_rmse ↓	
NDDR_CNN[3](D)	39.330 1	0.594 3	0.200 8	96.9×10^6
cross_stitch[36](D)	38.749 7	0.597 3	0.197 8	75.8×10^6
PADnet[6]	39.217 0	0.641 7	0.215 7	66.9×10^6
MTI_Net[7]	35.424 1	0.565 6	0.190 3	81.8×10^6
MTAN[2]	39.468 9	0.595 3	0.194 2	67.0×10^6
MTL_SPP(本文)	39.611 9	0.584 0	0.192 8	83.9×10^6
MTL_SSW(本文)	39.879 6	0.575 5	0.190 4	89.9×10^6

相较于基于预测结果的多任务框架 PADnet 和 MTI_Net，PADnet 在语义分割任务上的 MIOU 是 39.217%，比本文的 MTL_SSW 低 0.566%；深度估计任务的 rmse 是 0.641 7，比本文的 MTL_SPP 高0.057 7。MTI_Net 的语义分割 MIOU 为 35.424 1%，比本文 MTL_SPP 低 4.187 8%，比 MTL_SSW 低 4.363%，深度估计性能略好于本文提出的模型，但是两个任务的综合性能比本文方法更差。

本文提出的两个模型中，MTL_SSW 在 MTL_SPP 的基础上加入可选择权重融合模块，多层次融合任务特定层的特征，使得网络能充分学习两个任务的互补信息。MTL_SSW 语义分割的 MIOU 比 MTL_SPP 高 0.267 7%，rmse 和 log_rmse 分别比 MTL_SPP 低 0.008 5 和 0.002 4，在联合语义分割和深度估计两个任务上获得了最好的预测性能。

(2)在 SUNRGBD 数据集上的对比结果

在 SUNRGBD 数据集上的对比实验结果如表 2 所列。本文提出的 MTL_SPP 的 MIOU 比 MTAN 高 0.507 2%，rmse 和 log_rmse 比 MTAN 分别低 0.002 6 和 0.004 1；而 MTL_SSW 的 MIOU 比 MTAN 高 0.970 4%，rmse 和 log_rmse 比 MTAN 分别低 0.015 7 和 0.011。MTL_SSW 的 rmse 和 log_rmse 只比 MTI_Net 分别高 0.003 和 0.001 4，但是 MTL_SSW 的 MIOU 比 MTI_Net 高 2.471 8%。相比于现有联合语义分割和深度估计的多任务学习算法，本文算法在 SUNRGBD 数据集上均取得了较好的平均性能，其中 MTL_SSW 取得最优的预测结果。这表明本文算法在不同数据集上均可以学到对两个任务更具判别性的特征表示，能有效适应数据集中类别和场景的多样性。

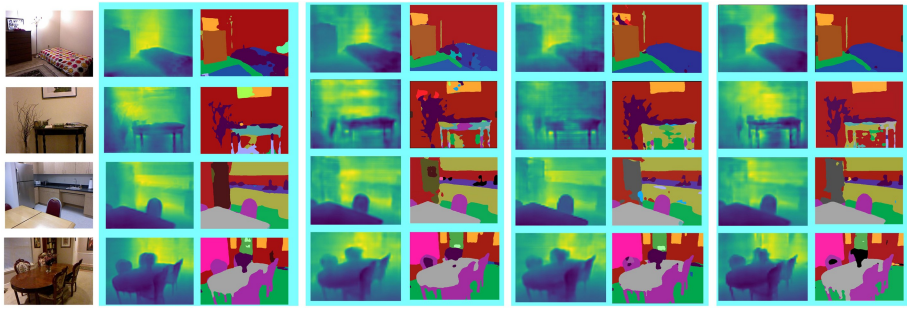
表 2　在 SUNRGBD 数据集上不同方法的实验对比

Table 2　Experimental comparison of different methods on SUNRGBD dataset

方法	语义分割	深度估计	
	MIOU ↑ /%	rmse ↓	log_rmse ↓
NDDR_CNN[3](D)	48.054 1	0.487 3	0.178 3
cross_stitch[36](D)	47.930 7	0.500 7	0.172 1
PADnet[6]	47.879 4	0.528 7	0.185 3
MTI_Net[7]	46.782 9	0.450 9	0.148 9
MTAN[2]	48.286 3	0.469 6	0.161 3
MTL_SPP(本文)	48.789 0	0.467 0	0.157 2
MTL_SSW(本文)	49.254 7	0.453 9	0.150 3

4.3　结果可视化

图 8 给出了 MTAN[2] 方法与 MTL_SPP 和 MTL_SSW 方法，在 NYUD_V2 数据集的几张经典图片上，深度估计和语义分割结果的可视化对比。图中每一种方法的左列是深度估计结果，右列是语义分割结果。可以看出，MTAN 算法对“墙”和“地板”等纹理简单的类，分割效果比较好，但是面对难以区分的类别时，比如第一行中的“格子床单”类和第二行的“画框”类，分割效果欠佳。它由于缺乏足够的语义辨别能力和对几何位置关系的学习，导致容易混淆的像素点被集中分类错误，不同物体的同深度值像素点估计不准确。相比于 MTAN，MTL_SPP 通过 SE 激励模块和金字塔池化模块能从共享特征中学习到互补信息，提升两个任务的性能，在各种复杂的室内场景都取得了精确的语义分割和深度估计结果，例如第一行中的“格子床单”第三行的“冰箱”和“柜子”等。



(a)原图　(b)MIT_Net　(c)MTAN　(d)MTL_SPP　(e)MTL_SSW

图 8　在 NYUD_v2 数据集上的语义分割和深度估计可视化结果对比

Fig. 8　Comparison of semantic segmentation and depth estimation visualization results on NYUD_v2 dataset

更进一步，本文提出的 MTL_SSW 在共享特征的基础上，通过 SW 模块进行不同任务的多级信息的相互融合和相互学习，获得了比 MTL_SPP 更好的性能。比如对第二行“树枝”的分叉以及第四行的小物体“花瓶”，MTL_SSW 得到了精确的分割效果；对第四行形状复杂的“桌子”和“凳子”，得到了准确的深度估计。

4.4 消融实验

本小节在 NYUD-v2 数据集上，对提出的两个多任务学习模型 MTL_SPP 和 MTL_SSW 进行了较详细的消融实验，以分析模型的不同组件对性能的影响。

4.4.1 MTL_SPP 消融分析

为了分析本文所提出的 MTL_SPP 中的 SE 模块和金字塔模块对多任务学习性能的影响，在 NYUD_v2 数据集上进行了消融实验。实验结果如表 3 所列，其中“STL_ASPP”表示在主干网络后加入空洞空间金字塔池化模块的单任务模型，“MTL”表示在主干网络后面直接加两个任务分支，“MTL_SE”表示把 MTL_SPP 中的特征金字塔池化模块移除，“MTL_PP”表示在“MTL”的基础上，在两任务分支上增加金字塔模块。

表 3 MTL_SPP 在 NYUD_v2 数据集上的消融实验结果

Table 3 Results of MTL_SPP ablation experiments on NYUD_v2 dataset

方法	语义分割	深度估计		参数
	MIOU ↑ /%	rmse ↓	log_rmse ↓	
STL_ASPP	39.2452	0.6000	0.2050	37.82×10^6
MTL	34.5325	0.6829	0.2326	53.18×10^6
MTL_SE	34.8917	0.6779	0.2280	55.11×10^6
MTL_PP	37.4387	0.6779	0.2364	66.43×10^6
MTL_SPP	**39.6119**	**0.5840**	**0.1928**	83.91×10^6

从表 3 的结果可以看出，MTL_SPP 获得的语义分割和深度估计分割性能均高于单任务基线 STL_ASPP，其 MIOU 比 STL 高 0.3667%，rmse 和 log_rmse 分别比 STL 低 0.02 和 0.0122。相比于多任务基线结构 MTL，MTL_SPP 的语义分割 MIOU 高 5.0794%，深度估计 rmse 和 log_rmse 分别低 0.0989 和 0.0398。只加入 SE 模块的“MTL_SE”和只加入金字塔池化模块的“MTL_PP”，其语义分割性能比 MTL_SPP 分别降低了 4.7202% 和 2.1732%，深度估计任务的 rmse 分别比 MTL_SPP 高 0.0939 和 0.0939，log_rmse 分别高了 0.0352 和 0.0436。这表明本文提出的 MTL_SPP 模型能有效结合 SE 模块特征选择学习作用，以及金字塔池化模块特征增强的作用，从而获得更精细的轮廓分割结果和深度估计结果。

4.4.2 特征增强模块分析

本文提出的两个多任务学习网络使用了不同的特征增强模块，MTL_SSW 框架采用了空洞空间金字塔池化模块，MTL_SPP 使用了金字塔池化模块。本实验将两个方法中的特征增强模块进行了互换，对比分析了两种特征增强模块在提出的两个多任务学习框架中的性能影响。实验结果如表 4 所列，表中“MTL_SE+ASPP”表示将 MTL_SPP 中的 PP 模块替换成 ASPP 模块，“MTL_SW+PP”表示将 MTL_SSW 中的 ASPP 模块替换成 PP 模块。从表 4 的实验结果可以看出，MTL_SPP 框架更加适合 PP 模块来增强特征表示能力。ASPP 模块更适合于多模态信息交互特征的增强，所以 MTL_SSW 使用它能获得更好的性能。

表 4 不同组合方式的性能对比

Table 4 Performance comparison of different combination methods

方法	语义分割	深度估计		参数
	MIOU ↑ /%	rmse ↓	log_rmse ↓	
MTL_SE+ASPP	37.9027	0.6219	0.2026	68.5×10^6
MTL_SW+PP	38.0569	0.6211	0.2067	103.2×10^6
MTL_SPP	**39.6119**	**0.5840**	**0.1928**	83.9×10^6
MTL_SSW	**39.8796**	**0.5755**	**0.1904**	89.9×10^6

4.4.3 MTL_SSW 消融分析

MTL_SSW 的 SW 模块通过一个可训练权重矩阵（见式(8)）来学习不同任务间信息的融合和交互方式。为了分析 SW 模块中可训练权重矩阵的参数初始化对模型训练效果的影响，本实验对 SW 层的参数矩阵初始化为值分别是 (α,β) 的对角矩阵，并使用不同的 (α,β) 值在 NYUD_v2 数据集上进行了实验，实验结果如表 5 所列。其中 (α,β) 表示语义分割任务的特征选择矩阵的初始化权重，(β,α) 为深度估计任务的特征选择矩阵的初始化权重。

表 5 可训练权重矩阵的参数初始化对 MTL_SSW 的影响

Table 5 Effect of parameter initialization of trainable weight matrix on MTL_SSW

初始化方法	语义分割	深度估计	
	MIOU ↑ /%	rmse ↓	log_rmse ↓
MTL_SSW(0.2,0.8)	37.9079	0.6086	0.2034
MTL_SSW(0.8,0.2)	38.4932	0.5995	0.2030
MTL_SSW(0.5,0.5)	37.5080	0.6095	0.2047
MTL_SSW(0.1,0.9)	37.6717	0.6014	0.1998
MTL_SSW(0.9,0.1)	**39.8796**	**0.5755**	**0.1904**

从表 5 的实验结果可以看出，在相同的训练环境下，在初始化权重为(0.9,0.1)，即对于语义分割子网络，初始化融合特征中语义分割特征的权重为 $\alpha=0.9$，深度估计特征的权重为 $\beta=0.1$，而对于深度估计子网络，初始化融合特征中深度估计特征的权重为 $\beta=0.9$，语义分割特征的权重为 $\alpha=0.1$ 时，在相同迭代次数情况下，可以获得最好的性能。这也说明在网络刚开始训练时，更多地利用自身任务的信息，以及少部分其他任务的信息作为辅助，能获得更好的训练效果。

结束语 本文提出了两种多任务学习网络结构 MTL_SPP 和 MTL_SSW，联合学习语义分割和深度估计任务，旨在发掘特定任务中隐藏的有效信息来提高最终的任务预测，平衡语义分割和深度估计任务的性能。基于压缩激励模块和金字塔池化的多任务学习框架 MTL_SPP，使用级联的 SE 模块作为任务特定的特征提取器，聚焦式学习共享特征池的特征，最后利用特征金字塔来增强特定任务特征的学习。基于压缩激励和可选择权重的多任务学习框架 MTL_SSW，利用 SW 模块融合任务特定的 SE 模块输出特征，让语义分割特征和深度估计特征相互指导和优化，从而学习到单个任务无法学习到的判别性特征，最后用空洞空间金字塔池化模块增强特征学习。在 NYUD-v2 数据集和 SUNRGBD 数据集上的实验结果表明，本文提出的模型可以有效处理复杂的室内场景，获得了很好的语义分割和深度估计性能。本文的方法具有良好的扩展性，下一步的工作将研究把本文提出的方法扩展到 3 个以上的多任务中。

参 考 文 献

[1] ZHANG Y,LIU J W,ZUO X. Survey of Multi-Task Learning[J]. Chinese Journal of Computers,2020,43(7):1340-1378.

[2] LIU S,JOHNS E,DAVISON A J. End-to-End Multi-Task Learning with Attention[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society,2019:1871-1880.

[3] GAO Y,MA J,ZHAO M,et al. NDDR-CNN:Layerwise Feature Fusing in Multi-Task CNN by Neural Discriminative Dimensionality Reduction[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Computer Society,2019:3205-3214.

[4] GAO Y,BAI H,JIE Z,et al. MTL-NAS:Task-Agnostic Neural Architecture Search towards General-Purpose Multi-Task Learning[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2020:11540-11549.

[5] WANG P,SHEN X,LIN Z,et al. Towards unified depth and semantic prediction from a single image[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society,2015:2800-2809.

[6] XU D,OUYANG W,WANG X,et al. PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society,2018:675-684.

[7] VANDENHENDE S,GEORGOULIS S,GOOL L V. MTI-Net: Multi-Scale Task Interaction Networks for Multi-Task Learning[C]// European Conference on Computer Vision. Piscataway: Springer,2020:527-543.

[8] ZHOU D,FANG J,SONG X,et al. Joint 3D Instance Segmentation and Object Detection for Autonomous Driving[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Computer Society,2020:1836-1846.

[9] ZHANG Z,FIDLER S,URTASUN R. Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016: 669-677.

[10] ZHANG H,LUO G,TIAN Y,et al. A Virtual-Real Interaction Approach to Object Instance Segmentation in Traffic Scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2021,22(2):863-875.

[11] KENDALL A,GAL Y,CIPOLLA R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2018: 7482-7491.

[12] LI X,WANG W,HU X,et al. Selective Kernel Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Computer Society,2019:510-519.

[13] EIGEN D,FERGUS R. Predicting Depth,Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture[C]// IEEE International Conference on Computer Vision. Piscataway:IEEE,2015:2650-2658.

[14] EIGEN D,PUHRSCH C,FERGUS R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network[C]// NIPS. 2014:2366-2374.

[15] KLINGNER M,TERMÖHLEN J A,MIKOLAJCZYK J,et al. Self-supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance[C]// ECCV. 2020: 582-600.

[16] YIN W,LIU Y,SHEN C,et al. Enforcing Geometric Constraints of Virtual Normal for Depth Prediction[C]// IEEE International Conference on Computer Vision. Piscataway: IEEE, 2019: 5683-5692.

[17] FU H,GONG M,WANG C,et al. Deep Ordinal Regression Network for Monocular Depth Estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 2002-2011.

[18] LIU J P,WEN J W,LIANG Y L. Monocular Image Depth Estimation Based on Multi-Scale Attention Oriented Network[J]. Journal of South China University of Technology(Natural Science Edition),2020,48(12):52-62.

[19] LONG J,SHELHAMER E,DARRELL T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.

[20] CHEN L-C,PAPANDREOU G,KOKKINOS I,et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018,40(4):834-848.

[21] RONNEBERGER O,FISCHER P,BROX T. U-Net:Convolutional Networks for Biomedical Image Segmentation[C]// MICCAI. Springer,2015,9351:234-241.

[22] CHEN L C,ZHU Y,PAPANDREOU G,et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[C]// ECCV. 2018:833-851.

[23] LIU W,RABINOVICH A,BERG A C. ParseNet:Looking Wider to See Better[C]// ICIL 2016. 2016.

[24] ZHEN M,WANG J,ZHOU L,et al. Learning Fully Dense Neural Networks for Image Semantic Segmentation[C]// AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019:9283-9290.

[25] CHEN L-C,PAPANDREOU G,SCHROFF F,et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. arXiv:1706.05587v3,2017.

[26] ZHAO H,SHI J,QI X,et al. Pyramid Scene Parsing Network[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Computer Society,2017:6230-6239.

[27] REN T C,HUANG X S,DING W L,et al. Global Bilateral Segmentation Network for Segmantic Segmentation[J]. Computer Science,2020,47(S01):5.

[28] YU C,WANG J,PENG C,et al. BiSeNet:Bilateral Segmentation Network for Real-time Semantic Segmentation[C]// European Conference on Computer Vision. 2018.

[29] ZHAO H,ZHANG Y,LIU S,et al. PSANet:Point-wise Spatial Attention Network for Scene Parsing[C]// European Confe-

rence on Computer Vision. Piscataway: Springer, 2018, 11213: 270-286.

[30] YANG J, DANG J S. Semantic segmentation of 3D point cloud based on contextual attention CNN[J]. Journal on Communications, 2020, 41(7): 195-203.

[31] DUAN L J, SUN Q C, QIAO Y H. Attention-Aware and Semantic-Aware Network for RGB-D Indoor Semantic Segmentation[J]. Chinese Journal of Computers, 2021, 44(2): 275-291.

[32] SEICHTER D, KÖHLER M, LEWANDOWSKI B, et al. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis [J]. arXiv: 2011. 06961v3, 2020.

[33] ZHANG Z, LUO P, LOY C C, et al. Facial Landmark Detection by Deep Multi-task Learning[C]// European Conference on Computer Vision. Piscataway: Springer, 2014, 8694: 94-108.

[34] RANJAN R, PATEL V M, CHELLAPPA R. HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2019, 41(1): 121-135.

[35] LU Y, KUMAR A, ZHAI S, et al. Fully-adaptive Feature Sharing in Multi-Task Networks with Applications in Person Attribute Classification[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2017: 1131-1140.

[36] MISRA I, SHRIVASTAVA A, GUPTA A, et al. Cross-stitch Networks for Multi-task Learning[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016: 3994-4003.

[37] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// IEEE International Conference on Computer Vision. Piscataway: IEEE 2017: 2980-2988.

[38] REN S, HE K, GIRSHICK R B, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[39] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016: 770-778.

LUO Huilan, born in 1974, Ph. D, professor. Her main research interests include machine learning and pattern recognition, multi-task learning, etc.