中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

硕 士 学 位 论 文

基于深度学习的视觉多任务密集预测算法研究

Research on Deep Learning Based Multi-Task Dense Prediction

论文作者 杨雨奇 指导教师 程明明 教授
申请学位 工学硕士 培养单位 南开大学
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 李重仪 教授 评 阅 人 匿名评阅

南开大学研究生院

二〇二五年三月

摘要

深度学习通过为不同的视觉任务训练对应的神经网络模型来解决不同的视觉问题。在同时需要多种密集视觉信息的应用任务场景下，多任务密集预测通过让一个模型在一次推理中解决多个密集预测任务，从而有效提升了效率。现有的多任务密集预测方法在模型结构和建模范式两方面存在局限：在模型结构方面，现有方法大多使用静态模型，无法针对不同的图像样本得到多样性的特征，因此无法提升任务专用特征的区分度，在建模范式方面，现有方法大多基于判别式方法，难以建模预测目标的条件概率分布，因此在细节预测上相比生成式方法有一定劣势。

针对模型结构上的局限，近年来的多任务密集预测模型开始采用混合专家模型的策略，但是其方法都因专家网络数量增加引入的更多参数与计算成本导致效率受限。受到低秩适应方法的启发，本文将专家网络中普通卷积的权重转化为低秩分解之后的形式。由于低秩专家拥有更少的参数，并且可以动态地参数化为通用卷积，减少模型的参数量和推理计算量。

针对建模范式方面的性能局限，本文提出一种将扩散模型引入多任务密集预测的方法。该方法在解码器中引入条件扩散过程，通过创新的联合去噪扩散机制捕捉任务间关联性，而非独立处理不同任务的噪声标签。利用联合去噪扩散机制，本文的方法能够利用扩散模型这一生成式模型更好地建模预测目标的条件概率分布并建立任务间的关系，从而提升不同任务的总体性能。

本文研究内容以及贡献可总结如下：1) 针对混合专家模型在效率上的局限，本文提出一种将专家网络中普通卷积的权重转化为低秩分解之后的形式的方法。这一方法通过为专家网络增加低秩限制，从而有效提升了混合专家模型的效率。2) 针对判别式方法难以建模预测目标的条件概率分布的局限，本文将扩散模型引入多任务密集预测，并提出创新的联合去噪机制，从而能够更好地建模任务间关系并提升多项任务的总体性能。3) 本文在两个常用的多任务密集预测数据集和共六项不同的密集预测任务上进行了广泛的实验，实验结果表明，本文提出的低秩专家模型和联合去噪策略能够显著提升模型在所有任务上的表现。

关键词： 多任务密集预测；混合专家模型；低秩结构；扩散模型；多任务学习

Abstract

In application scenarios requiring multiple dense visual information, multi-task dense prediction enhances efficiency by enabling a single model to address multiple dense prediction tasks in one inference. The previous multi-task dense prediction methods had limitations in the aspects of model architectures and modeling paradigms. For the model architectures, the previous methods leveraged static architectures, which fail to capture diverse features for different image samples, reducing the distinctiveness of task-specific features. For the modeling paradigm, the previous methods were based on discriminative methods, which struggle to model the underlying conditional distribution of target tasks, and perform less effective in detail compared to generative methods.

To address the limitations of static models, recent approaches have adopted mixture of experts strategies. However, these methods face efficiency challenges due to increased parameters and computational costs with more expert networks. Inspired by low-rank adaptation, this thesis leverages the low-rank form of standard convolutions in expert networks. Low-rank experts, with fewer parameters, can be dynamically parameterized as general convolutions, reducing model size and computation cost.

To overcome the performance limitations of discriminative methods, this thesis introduces diffusion model into multi-task dense prediction. The method incorporates a conditional diffusion process in the decoder, using an innovative joint denoising diffusion process to capture tasks relations instead of independently processing noisy labels for different tasks. The proposed joint diffusion process allows better modeling of the underlying conditional distribution of prediction targets and tasks relations, enhancing overall performance across different tasks.

The contributions of this thesis can be summarized as follows: 1) To address the efficiency limitations of mixture-of-expert models, we propose using the low-rank form of standard convolutions in expert networks, effectively improving the efficiency by imposing low-rank constraints. 2) To tackle the challenge of modeling underlying conditional distribution with discriminative methods, we introduce diffusion models to multi-

task dense prediction and propose a novel joint denoising diffusion process, better capturing task relations and boosting overall performance. 3) Extensive experiments on two common multi-task dense prediction datasets across six different tasks demonstrate that our proposed mixture-of-low-rank-expert model and joint denoising diffusion process significantly enhance model performance across all tasks.

Key Words: multi-task dense prediction; mixture of expert; low-rank architectures; diffusion model; multi-task learning

目录

第一章 绪论

第一节 研究背景和意义

近年来，深度学习通过构建多层神经网络来解决各种复杂的感知问题。在深度学习技术的驱动下，计算机视觉领域的许多任务都取得了巨大的进展，尤其是图像分类[1]，语义分割[2–5]，目标检测[6]，深度估计[7, 8] 等。在这些任务中，有一类任务需要对图像中的每一个像素都进行预测，被称为密集预测任务，如语义分割和深度估计。这些任务能够为图像提供更加细致的感知，在许多实际领域都有着重要的应用，如自动驾驶[9–11] 和医学诊断[12] 等。在实际应用的场景中，模型往往会需要图像不同方面的信息，所以需要同时对一张图像执行多个感知任务，如在自动驾驶的场景下，会同时需要摄像机拍下图像的语义信息判断路上是否有行人以及深度信息判断行人与车之间的距离。若为每个感知任务单独配置神经网络模型，并分别让这些模型执行各自的推理任务，则会在内存和计算量两个方面都产生一定的负担。考虑到许多实际应用场景部署在内存和计算资源受限的边缘设备上，因此在内存和计算量上的负担都会对模型的部署产生不利的影响。针对这一问题，现有的技术将不同的任务集成到一个模型上，让一个模型在一次推理中同时解决多个任务，有效地缓解了原有配置方式在效率上的局限性。

在将不同任务的模型集成到单一模型上的时候，通常会在此模型上通过多任务的监督信号学习能够为多个任务所共享的特征，并将共享的特征输入任务特定的网络支路得到任务特定的特征，同时在任务特定的特征之间学习任务之间的联系。为了提升多任务模型的性能，往往需要对这三项要素的建模进行改进。对于任务共享特征而言，通常会考虑如何共享模型的参数来学习任务之间共享的特征，并在模型的不同层设计不同的方式进行参数的共享[13, 14]。而对于任务特定特征而言，研究者会希望能够让不同任务所对应的特征能够更有区分度[15]。最后，相比起单任务模型，多任务模型能够通过共享任务之间互补的信息，或者让不同任务充当彼此的正则化从而提升整体的任务性能[16]，因此建立任务间的关系也能够提升多任务模型的总体性能。具体到多任务密集预测这一

任务上的时候，研究者通常会改进模型的编码器或者解码器来更好地建模上述的三种特征。编码器用于对图像进行特征编码，在这一阶段得到的特征往往是为多个任务所共用的，因此其主要用于建模任务共享特征。而解码器用于将编码器得到的特征解码为具体的任务预测，在这一阶段需要把任务共享的特征解码为各个任务专有的特征。同时，由于此时的特征更能反应不同任务，所以对任务间关系的建模也往往在这一阶段进行。

以往基于解码器结构改进的方法[17–19] 常通过手工设计的解码器结构来构建任务特定特征与任务之间的关系，但是这样的静态网络结构并非最优。一方面，静态的网络结构因为无法根据不同样本而改变结构，所以其生成特征的多样性相比动态的网络结构要更差，而这一差距会影响不同任务专有特征之间的区分度，最终导致预测质量的下降。另一方面，对于不同的样本而言，其任务特定特征之间的联系也应当是不同的，但是静态的网络结构往往使用手工设计或者神经网络结构搜索[20] 的方式对所有的样本建立同样的参数共享模式，因此也不利于对不同样本建立其专有的任务间联系。

为了突破静态解码器结构的上述缺陷，近来有一些方法[15] 将混合专家模型这一动态的网络设计思路融入解码器设计中。混合专家模型通常包括多个结构相同或类似的专家网络与一个路由网络，在新样本的特征输入混合专家模型的模块时，路由网络会根据这一特征计算出选择不同的专家网络处理这一特征的概率。随后，模型选取其中前 k 大概率的专家网络来处理这一特征，并把处理后的特征以路由网络预测的概率值作为权重进行加和，作为这一样本的特征。在多任务密集预测中，不同任务会对应不同的路由网络，因此可以对不同的样本在不同的任务中选取不同的专家网络，一方面可以生成更加具有多样性的特征，另一方面也可以动态地建立任务间关系。本文提出的低秩混合专家模型在混合专家模型的基础上，进一步改进了专家网络的结构与路由方式，从而使其在性能与效率两个方面都取得了新的突破。

在结构上的缺陷之外，密集预测任务模型的传统建模范式也阻碍了多任务密集预测这一领域的发展。以往的密集预测任务往往利用判别式模型对任务的分布进行建模，这一训练思路让模型对输入图像的每一个像素进行预测，计算每一个像素代表的概率分布或者回归值，而却忽略了对于预测结果这一掩码表征的内在条件分布[21]，因此在细节上的预测会有一定劣势。与基于判别式的密集预测方法相对，基于生成式的密集预测方法可以将不同的预测任务看成是在

输入图像条件下的预测掩码生成。这一范式可以捕获预测掩码的内在条件分布，更好地构建其在细节上的预测。

基于此原因，近来很多密集预测框架[22, 23] 都以生成式方法作为其基础。其中，大部分的方法利用扩散模型[24, 25] 来解决各种不同的密集预测任务。作为目前最为流行的生成式方法框架，扩散模型以扩散过程为基础学习生成目标的分布。扩散过程是指在一张图像上逐步增加高斯噪声，最后得到一张纯高斯噪声图像。而扩散模型通过训练一个去噪模型，对纯高斯噪声图像进行迭代去噪，模拟扩散过程的逆过程，从而从纯高斯噪声图像中生成原本的图像。目前，扩散模型在图像和视频生成领域有着非常卓越的效果[26]。因此，许多方法也考虑将扩散模型强大的性能用于密集预测任务上。在密集预测任务中，和图像生成任务有一定不同，其需要学习的是一个条件分布，而条件则为给定的输入图像的特征。同时，其去噪目标也不再是图像本身，而是图像所对应密集预测任务的预测掩码。在经过这两项修改后，扩散模型可以在各种密集预测任务，如语义分割[22] 和深度估计[23] 中，对纯高斯噪声进行迭代去噪，并最后得到对应任务的预测掩码。

此外，基于判别式方法的密集预测模型虽然在模型大致结构上基本相似，但还是因为任务内在的特性的不同而让模型在具体设计上有着区别，这一点在构建密集预测模型的通用模型以及多任务模型时都会产生阻碍，而基于生成式的密集预测方法往往可以更好地统一不同的密集预测任务。如在前文所述的扩散模型的框架下，不同的密集预测任务可以被统一为对纯高斯噪声的迭代去噪，因此其在构建通用的密集预测模型上也会更有优势。近年来，一些方法[27] 通过把密集预测任务的编码器-解码器结构与扩散模型的原理相结合，先利用编码器将图像进行编码得到图像的编码特征，然后利用基于扩散模型的解码器以图像的编码特征为条件对高斯噪声进行迭代去噪，从而得到最后的预测结果。这一通用框架有效地统合起了不同的密集预测任务，同时也通过分离编码器和解码器减少了扩散模型的迭代去噪过程所增加的计算负担，在利用了扩散模型的强大性能的同时也大大地提升了模型的效率。

尽管扩散模型在通用的密集预测模型上的有效性已经得到了验证，但是基于扩散模型的多任务密集预测模型还亟待探索。值得注意的是，将扩散模型应用于多任务模型上的时候并不能简单地通过方法的迁移来完成，而是存在至少两方面的挑战。一方面，因为扩散模型在推理时需要多次迭代前向过程进行去

噪，所以其计算量相比起一般的框架会更高。尽管存在如隐去噪扩散模型[28] 等加速去噪过程的方法，但是多次迭代所带来计算量负担仍然是不可忽略的。同时，多任务模型往往会在解码器上设置多条任务特定支路，因此如果是由基于扩散模型的解码器来构建每条任务特定支路的话，那么其增加的计算负担会随着任务数量的增多而成倍放大，这在重视效率的多任务领域是不可接受的。由此可看出，现有的方法需要更加高效的方法来结合多任务密集预测模型与扩散模型。另一方面，目前用于密集预测任务的扩散模型方法往往注重于单一任务的性能，而忽略了多个任务之间的关系。但是，如果无法很好地建模任务与任务间关系，那么多任务密集预测模型的性能也会大幅下降。因此，如何利用扩散模型方法强大的强大的能力来捕获跨任务关系的潜力也是十分重要的。针对上述的两个挑战，本文提出了一个新颖的基于扩散模型的多任务密集预测框架。在这个框架中，本文用联合去噪的方式将多个任务的迭代去噪过程结合在一起，在一次迭代去噪中得到不同任务的结果，保证了多任务模型的效率。同时，本文还在去噪的过程中显式地建模了任务与任务之间的关系，提升了模型在不同任务上的整体性能。

最后，本文前述的两个方向都对提升多任务密集预测方法的性能有着重要作用，同时，在这两个方向上的改进也能够达到相互促进的效果，从而为模型的性能带来进一步的提升。

第二节　国内外研究现状

一、　多任务密集预测方法中网络结构的研究现状

许多多任务密集预测方法都将其重点放在如何设计网络结构，从而不同的任务之间能够通过适当的参数共享来建立任务与任务之间的关系，并提升任务的整体性能。如有些方法[13] 专注于在解码器阶段为不同的任务设计不同的支路，并且在特定的层添加特征交互模块，从而让不同任务的特征与其他任务的特征之间形成互补，促进不同任务性能的提升。除了性能之外，效率也是多任务密集预测方法需要考虑的因素，因此也有另外一些方法[18–20, 29] 让不同任务之间共享大部分参数，从而减少模型的参数量与计算成本。具体来说，这类方法会将编码器进行共享，并在解码器阶段为不同任务设置多条支路，同时设计特殊的模块来建模任务与任务之间的联系。同时，这类方法往往会借助解码的初步结果，获得更加具有区分度的任务特定特征，从而利用特征本身[18]，不同

像素上特征的相似度[20]，不同特征间的注意力图[19]，或者不同尺度的特征[17]，特征上不同感受野的上下文[30]来建立任务之间的关系。

因为上述的方法往往局限于手工设计的模型架构，所以后来的方法在编码器阶段[31, 32]或者解码器阶段[30]利用网络结构搜索的技术，在树状搜索空间中[31, 32]进行搜索，并通过端到端的优化找到一个同时在性能和效率上都具有优势的模型架构，或者用来确定使用哪种类型的上下文来建模任务之间的关系[30]。这些方法搜寻到的网络架构往往比手工设计的网络架构要更加有效，因此可以更好地建模任务与任务间的关系。但是，这些方法所找到的网络结构和手工设计的结构同样是静态的，无法根据输入样本的不同而进行变化，所以会导致生成特征的多样性受限，同时不同任务对应的任务特定特征的区分度也会较低。针对这一局限，许多多任务密集预测方法[15, 33–35]也开始将以混合专家模型[36, 37]为首的动态网络结构引入这一领域。混合专家模型包括多个专家网络和一个路由网络，其中专家网络用于对输入样本进行不同方面的特征提取，并由路由网络决定每个专家对最终结果贡献的概率。通常混合专家模型会使用稀疏的专家网络选择方法，选择概率前 k 大的专家的输出组成最终结果。通过此方式，不同样本会因为选择不同的专家网络来处理，从而实现了动态的模型结构。在多任务密集预测领域，往往会为不同的任务设置不同的路由网络，从而让不同任务根据样本动态地共享或者专属某个专家网络，建模更加多样化的任务间关系。现有的方法将混合专家模型作为编码器[33–35]的基本组成模块，或者作为解码器中建立任务间关系、增强任务专属特征区分度的特定模块[15]。但是，目前基于混合专家模型的多任务密集预测方法存在两个局限：第一个局限是无法解决多个专家带来存储与计算上的负担，因此往往选择减少专家的数量，从而无法充分利用多专家带来的性能提升；第二个局限是对于混合专家模型在构建多任务之间关系的作用上研究较少，没有将其路由模式基于多任务密集预测的特点进行合适的改进。

二、 多任务密集预测方法中建模范式的研究现状

现有的多任务密集预测方法大多以判别式的方法进行训练，将密集预测任务看作是一个逐像素的分类任务或者回归任务。这种方式虽然有效，但是忽视了对于预测结果这一掩码内在分布的建模，因此在细节部分的预测会有一定劣势。此外，在判别式的训练方法下，不同任务因为其损失函数不同，所以更新梯度会在大小和方向上都有较大的差异，这一点也会影响多任务密集预测的性

能。事实上，有许多较早期的多任务密集预测方法专注于平衡不同任务之间的梯度[38–42]，防止单一任务的梯度阻碍了其他任务的学习。同时，异质性的损失函数也会带来梯度冲突，早期的方法[42] 同样在这方面进行了研究，并取得了一定的效果。

因为判别式方法的上述局限，近来许多密集预测任务方法和构建通用模型的密集预测任务方法[23, 27, 43–45] 都开始尝试使用生成式的方式进行训练。目前，最强大的生成式方法为扩散模型[25, 28]，通过训练一个去噪网络，然后对高斯噪声进行迭代去噪，从而得到想要生成的目标。当运用于密集预测任务时，生成的目标会设置为对应任务的预测掩码，如对于深度估计来说，生成目标会设定为单通道深度图像[44]，而对于图像分割来说，生成目标会设定为多通道的掩码，其中通道数和生成的类别数相同[23]。此外，因为生成的预测掩码需要与输入图像对应，因此需要将图像编码作为扩散模型的条件，并计算生成目标的条件概率分布。通过此方式，扩散模型能够捕获对应任务的内在分布，同时还能够把不同的任务统一为图像的去噪任务，为建立通用的密集预测模型建立了基础。举例来说，Ji 等人[27] 将多个密集预测任务统一到了一个编码器-解码器的扩散模型框架中，其中编码器用于编码条件特征，解码器用于生成目标掩码。该方法还将解码器轻量化，从而减少迭代的去噪过程对模型推理产生的计算负担，提升了模型的效率。不过，现有的基于生成式方法的密集预测方法大多局限在单任务场景，而其在多任务密集预测中的表现还有待进一步的挖掘。

第三节　本文研究内容

多任务密集预测能够帮助场景理解模型部署到边缘设备上，因此具有重要的应用和研究价值。本文主要从模型架构和建模范式两个方面对以往的多任务密集预测方法进行改进。本文的主要贡献如 图 1.1所示。

在模型架构方面，本文从前沿的动态神经网络架构：混合专家模型入手。虽然现有的基于混合专家模型的多任务密集预测方法取得了很好的性能，但是由于引入多个专家网络而导致的存储与计算上的负担问题并没有被很好地解决。同时，这些方法缺乏对路由方式的进一步研究，忽视了明确建模所有任务之间全局关系的重要性。

针对这两方面问题，本文提出了一种新颖的多任务密集预测解码器结构，称为低秩混合专家模型。为了控制专家网络数量增加所带来的参数和计算成本，本

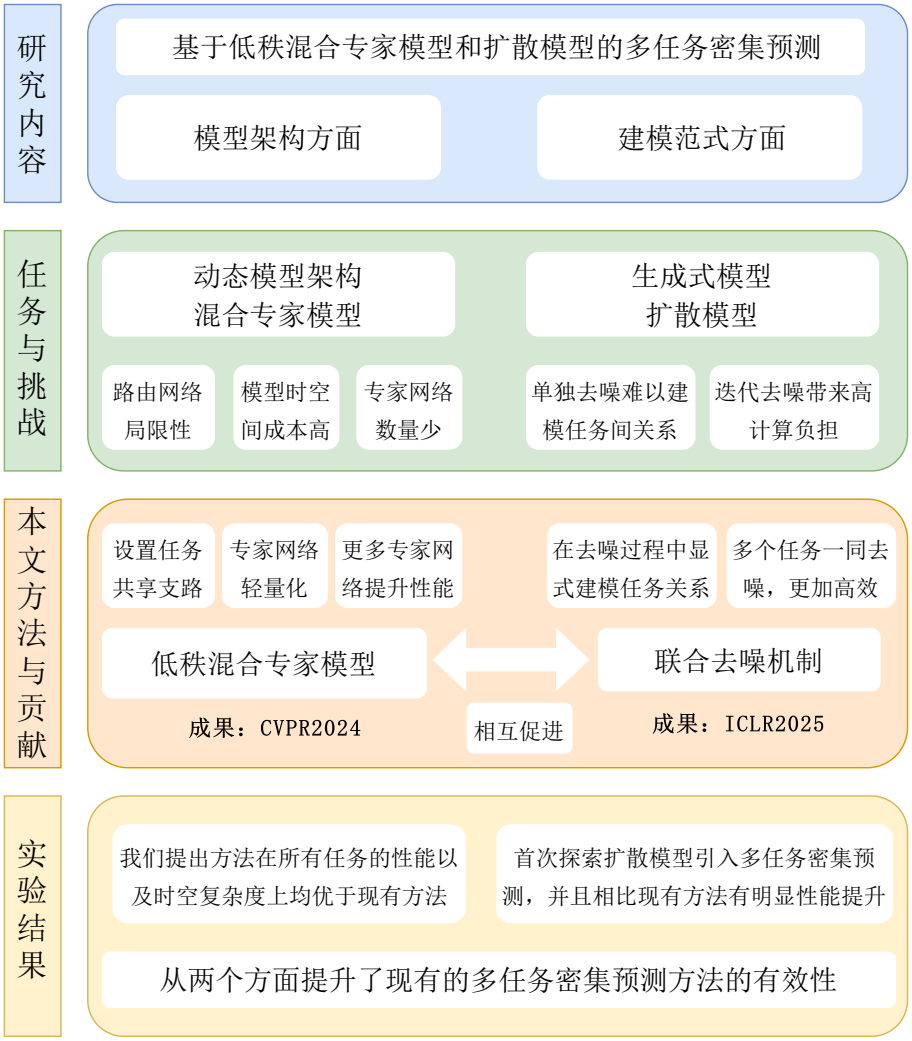

图 1.1 本文研究内容。

文从低秩适应获得灵感，提出对专家网络的参数矩阵进行显式的低秩约束，从而将其参数轻量化。受益于这一设计，本文所提出方法能够比现有的方法使用更多的专家数量及更大的感受野，以扩大模型的容量和表征能力，进一步发掘混合专家模型在多任务密集预测中的潜力。由于低秩专家拥有更少的参数，并且可以动态地重参数化为通用卷积，因此即使专家数量增加，参数量和计算成本也不会有很大变化。

此外，为了建模任务间的全局关系，低秩混合专家模型在混合专家模型的结构中增加了一个平行的通用卷积路径，每个任务特征都可以通过该路径进行显式参数共享。通过这种方式，本文利用了共享的参数来建立了所有任务之间的全局关系，大幅提升了任务的整体性能。通过对于混合专家模型所存在问题

的针对性改善，本文所提出的模型能够在性能和效率两方面都超越现有的方法。

在建模范式方面，因为现有的多任务密集预测方法基本基于判别式的建模范式，所以无法显式建模预测掩码的内在分布，导致其在细节上的预测效果较差。此外，扩散模型作为目前最为前沿的生成式方法，在多个领域展现出卓越的性能，但是现有的研究却鲜少探索其在多任务密集预测之中的作用。基于这两点，本文将扩散模型引入多任务密集预测任务中，并提出一种新型的多任务密集预测方法——TaskDiffusion，从而释放扩散模型在多任务密集预测中的潜力。TaskDiffusion 在解码器中引入条件扩散过程，而为了在去噪过程中捕获任务间的关系以提升性能，本文提出了联合去噪扩散机制。这一机制通过将多个任务的去噪过程统一到一个去噪过程里，同样也缓解了迭代去噪在多任务学习中产生的严重计算负担。

具体来说，本文首先将任务专属标签编码至统一的任务联合特征空间，从而消除冗余的任务专属编码过程。同时，本文还提出以任务专属多级特征作为条件的跨任务扩散编码器，显式建模跨任务与跨层级的交互关系，有效地提升了模型在所有任务上的性能。通过这种方式，本文将扩散模型引入多任务密集预测中，并利用其强大的能力捕获任务内以及任务间关系的内在分布，同时还利用联合去噪扩散机制缓解了迭代去噪带来的效率问题，为扩散模型在感知领域的进一步应用贡献了新的方案。

最后，如图中所示，本文所提出的两个方法从两个不同的方向对多任务密集预测方法的性能进行提升，并且具有相互促进的效果。当两个方法结合到一起时，能够实现比单个方法本身更强的性能。

第四节　本文结构

本文基于低秩混合专家模型和扩散模型两种方式提出了新的多任务密集预测方法，从而对现有方法在模型架构和建模范式两个方面的局限做出了针对性的改进，而本文也将会在这两个方面进行展开。具体来说，本文的正文部分总共有五章，其主要内容如下：

- 第一章主要介绍了多任务密集预测任务的目标和其意义，具体来说，本文从模型架构和建模范式两个方面详细介绍了多任务密集预测的研究背景，随后讨论了该任务在这两方面的研究现状，最后对本文的主要研究内容和贡献进行了简要的概括。

- 第二章主要介绍了与本文相关的工作，主要回顾了现有的多任务密集预测方法的发展现状，介绍了如混合专家模型等前沿模型结构相关的工作，以及扩散模型这一生成式方法在感知领域的发展。
- 第三章详细介绍了基于低秩混合专家模型的多任务密集预测方法，针对现有基于混合专家模型的多任务密集预测方法在效率和性能两方面上的局限，本文从低秩适应中得到灵感，通过对专家网络进行显式的低秩约束对模型进行轻量化，此外，本文还引入任务共享支路，从而在混合专家模型中建立所有任务之间的关系。本章第一小节介绍了研究动机以及其贡献，即现有的基于混合专家模型的多任务密集预测方法的局限性，而为了突破这些局限，本章在第二小节中具体阐释了本文所提出的低秩混合专家模型，并在第三小节中用定量和定性的实验证明其性能的优越。此外，本小节还对模型的各个组成部分和重要参数进行了消融实验，进一步地验证所提出方法的有效性。
- 第四章详细介绍了基于扩散模型的多任务密集预测方法，针对现有的多任务密集预测方法在建模范式上的局限，本文引入了目前生成式方法中最为前沿的扩散模型，同时提出了联合去噪扩散机制，从而在去噪过程中显式建立任务与任务之间的联系，同时缓解了迭代去噪在多任务学习中产生的严重计算负担。本章第一小节主要介绍在多任务密集预测中引入扩散模型的背景，从扩散模型在感知任务中的应用开始，介绍将扩散模型应用于多任务密集预测的挑战。针对这些挑战，本章在第二小节中具体介绍了本文所提出的 TaskDiffusion 方法，并且分析了其在效率上的优势。最后，本文在第三小节中用各种实验证明了本文所提出方法的有效性，并且用广泛的消融实验和可视化结果深入理解所提出方法有效的原因。
- 第五章总结了本文中提出的两项方法，并且从模型架构和建模范式两个方面入手，对多任务密集预测这一领域未来的发展进行展望。

第二章 相关工作

第一节 密集预测

在计算机视觉领域，密集预测任务指代那些需要对图像中的每一个像素都进行预测的任务。举例而言，如语义分割[2, 3, 46–58]，单目深度估计[59–64]，显著性物体检测[65–68]，法线估计[59, 69] 和边缘检测[70, 71] 等。这些任务需要在更精细的粒度上建模像素与像素之间的关系，同时还需要高层次的语义信息[3, 46] 为像素进行预测，因此比图像分类等任务更具有挑战性。通常来说，一个典型的密集预测任务方法会包含一个编码器和一个解码器，编码器负责将输入图像编码为高层次特征，而解码器负责将特征重新解码为目标预测掩码，解码器在解码时往往会利用编码器多层次的信息，从而同时捕获低层次的像素关系和高层次的语义信息。高层次特征和低层次特征之间的配合是密集预测方法的重要部分，举例而言，在语义分割[3] 和深度估计[72] 都有利用图像的条件随机场来细化最终预测结果，而条件随机场利用的则是近邻像素或超像素的距离这一低层次特征。早期的密集预测方法大多基于卷积神经网络[2, 3, 46–51]，但因为卷积神经网络的感受野有限，同时图像的分辨率会随着层数的增加而下降等问题，所以有一定局限性。在语义分割领域，一些经典的方法[46, 50] 通过在编码器中设置池化层或者带孔卷积等模型结构来缓解感受野的局限，并且都取得了一定的效果。近年来，随着视觉 Transformer[73] 这一强大的神经网络架构的提出，许多密集预测方法[52–55] 便基于视觉 Transformer 开发并且取得了卓越的效果。相比起卷积神经网络，视觉 Transformer 一方面可以借助自注意力机制构建全局的感受野，另一方面能够在神经网络的所有层保持相同的分辨率，因此更加适合用于密集预测任务。同时，一些密集预测方法[54, 55] 还借助自注意力机制，学习多个特殊标记，每个标记代表一个目标掩码，从而统一了语义分割、实例分割和全景分割三个不同的分割任务。目前，最为前沿的密集预测方法利用高参数量的视觉 Transformer 构建基础密集预测模型[74–76]，通过数据引擎[74] 和自监督学习[76] 弥补标注困难的问题，将密集预测任务的零样本场景下性能推向了新的境界。

第二节 密集预测中的多任务学习

在计算机视觉领域，针对密集预测任务的多任务学习已经得到了广泛的研究。

现有的方法可以分为两类，包括基于优化的方法和基于架构的方法[16]。优化方法[38–42] 的主要思路是通过平衡不同任务之间的梯度在模型上的更新，从而让模型能够在训练中同时学习到不同任务所需要的知识。如 Guo 等人[40] 根据任务训练的学习困难程度动态调整不同任务损失函数的权重，从而优先学习困难的任务。此外，Chen 等人[42] 针对多任务模型训练中不同任务梯度会具有相反的更新方向的问题，提出了一个基于概率的梯度掩码机制，从而减少不同任务之间梯度的冲突。

基于架构的方法旨在通过设计不同的神经网络架构，从而让单一模型能够进行多任务学习。更具体来说，基于架构的方法可以根据任务信息交互所在环节进一步分为两类，即以编码器为中心的方法和以解码器为中心的方法[16]。其中，以编码器为中心的方法[32, 77–79] 在编码器阶段重点进行任务之间信息的共享。Misra 等人[13] 提出了一种交叉缝合模块，用于让不同任务支路之间进行特征的共享与交互，从而使两个任务进行互补并提升两者的性能。Liu 等人[80] 利用软注意力机制让不同的任务支路从一个共享的骨干网络中选择其需要的特征，在建立了所有任务间共享的关系的同时还改善了交叉缝合网络难以提升任务数量的局限。在这些手工设计模型架构的方法之外，也有方法[31, 79] 利用神经网络结构搜索的方式来端到端地优化优化网络结构，从而能够更好地建模任务与任务间的交互。还有些方法[33, 34] 运用动态的网络结构来进一步提升编码器特征的多样性。

以解码器为中心的方法[15, 17–20, 29, 30, 81, 82] 不仅在编码器阶段通过共享参数来建立任务与任务之间的关系，在解码阶段也会设计精细的预测头来提取每个任务的任务特定特征并且设计模块来捕获跨任务关系。

因此，在以解码器为中心的方法中，如何设计网络结构来构建任务与任务之间的关系仍然是性能提升的重点。Xu 等人[18] 利用空间注意力机制，将初步预测的不同任务特征通过空间注意力加权求和，从而得到目标任务的特征，建立了像素级别的任务间关系。Zhang 等人[20] 更进一步，利用不同任务特征的像素相似度而非特征本身来建模不同任务之间的关系。之后，Vandenhende 等人[17] 将任务之间关系的发掘扩展到神经网络的不同尺度上，从而能够在不同的感受

野上建模任务间关系。Bruggmemann 等人[30] 使用了不同尺度下的上下文这一更加细致的特征作为建模的基础，同时使用注意力机制来建模任务间关系。同样，也有方法[30, 83] 将神经网络结构搜索引入编码器中心的方法，方便得到更优的任务间关系建模方式。相比起以编码器为中心的方法，以解码器为中心的方法的优势之一在于它们可以从现成的强大视觉骨干网络[73, 84, 85] 中受益，而不需要去从零训练不同任务专属的神经网络。如 Ye 等人[19] 首次将视觉 Transformer[73] 作为骨干网络引入到多任务密集预测领域，并且利用自注意力机制设计了一个多层次的解码器，实现了卓越的性能。最后，也有方法探索如何超越编码器为中心与解码器为中心这样的区分，设计一种在编码器与解码器中都能显式建模任务间关系的机制，如 Ye 等人[81] 通过为每个任务设置提示词标记，这些提示词标记在编码器中学习任务间与任务内关系，并在解码器中利用其建模任务共享特征、任务特定特征和任务间的特征交互。本文中提出的两种方法均属以解码器为中心的类别，研究如何利用混合专家模型技术有效地产生任务特定特征以及利用扩散模型更好地建模任务间关系。

第三节　混合专家模型

在以往的神经网络方法中，神经网络的结构大多为静态，不会根据不同的样本而改变。这样的一种方式对于增强神经网络的灵活性以及其生成特征的多样性来说是十分不利的。近年来，许多新颖的动态结构神经网络被提出，其中混合专家模型[36, 37] 取得了最为突出的效果，也被运用于如大语言模型[86, 87] 等多个热门领域中。具体来说，混合专家模型学习多个专家网络和一个路由网络。在推理时，模型利用训练好的路由网络来控制每个专家对最终输出贡献的概率。常用的混合专家模型往往采用稀疏的专家选择策略，即只选择概率最大的前 k 个专家网络的输出。稀疏的专家选择策略一方面减少了模型推理时的计算负担，另一方面也促使不同的专家去学习不同的知识，从而提升特征的多样性。在特征的多样性之外，混合专家模型还具有很强的缩放性[88]，即专家数量的提升往往会带来性能上的提升，这一特性也被大语言模型等追求巨大模型规模和强大性能的方法所青睐[86, 87]。最后，混合专家模型灵活的多专家结构也让其在一些特殊的领域得以发挥，如持续学习领域[89, 90] 和领域适应[91]。

具体来说，Le[90] 把前缀微调看作是往混合专家模型中增加新专家，从而通过设计一种特殊的路由网络来增加任务适应的效率。此外，Zhang 等人[91] 把不

同任务训练的适应器网络看作是不同的专家网络，并通过训练路由网络来重组这些专家网络用于适应新的任务，充分体现了专家混合模型的灵活性。

在上述的优势之外，因为混合专家模型还可以通过为不同任务配置不同路由网络的方法来动态地建模任务与任务之间的关系，所以也被广泛地用于多任务密集预测中。

现有的基于混合专家模型的多任务密集预测方法[15, 33–35] 主要是以编码器为中心的方法。如 Fan 等人[33] 设计了一个利用混合专家模型作为基本模块的视觉 Transformer，并且通过双缓存机制来实现高效的推理。Chen 等人[35] 通过自适应地扩张不同的路由网络选择专家网络的数量，从而能够让单一模型适应多个相互之间差距较大的任务。最近，Ye 等人[15] 首次将混合专家模型技术引入了以解码器为中心的任务。他们利用空间上下文感知门将来自不同专家网络的每个像素的特征进行结合。

这些方法将骨干网络特征分解为多个通用特征空间，并从中组合出判别性的任务特定特征。与上述基于混合专家模型的多任务学习方法不同，本文提出的方法首次在混合专家模型结构中明确构建了所有任务之间的全局关系，而不是通过任务特定的路由器隐式地完成这一工作。此外，本文中提出的低秩专家使得混合专家模型在效率上优于普通的混合专家模型结构，并且随着专家数量的增加，这种差距会逐渐扩大。

第四节 低秩结构

低秩结构因其高效性在深度学习中被广泛使用[92–95]。近年来，许多参数高效适应的方法[96–99] 都利用了低秩结构的特性对可更新参数进行轻量化。其中影响力最大的工作来自 Hu 等人[96]，他们为可更新参数增加显式的低秩约束，并利用参数矩阵的低秩分解完成了轻量化。这一想法的灵感来源于 Aghajanyan 等人[100] 在研究中的发现，即预训练模型和适应模型之间的权重差异仅仅依赖于一个低秩的矩阵。因此，通过学习一个额外的低秩矩阵，而不是调整整个层的参数矩阵，就足以让网络适应到一个新的任务上，换言之，适用于不同任务的模型在神经网络参数上的差异可以是低秩的。与本文的工作更相关的是，早期的多任务学习方法[94, 95] 利用低秩结构来建模任务通用特征，并通过线性组合生成任务特定特征。与这些方法不同，本文中提出的方法利用低秩结构来控制在混合专家模型中增加专家数量时的计算成本。

第五节 扩散模型

作为最前沿的图像生成方法，扩散模型与基于分数的生成模型[24, 25, 28] 比起之前的生成方法具有更强的特征多样性与稳定性，因此能够生成质量更好的图像。具体来说，扩散模型基于扩散过程的逆过程，利用训练的去噪模型从高斯噪声中进行迭代去噪，从而一步步生成需要生成的目标图像。尽管其性能突出，但因为依赖于迭代去噪，所以其效率相比其他的生成方法有一定局限。针对这一局限，许多加速扩散模型的方法被提出，如隐去噪扩散模型[28] 通过设计非马尔可夫的扩散过程，从而让模型从大量的去噪步中只需要采样特定的几步就可以完成最终的生成，大大提升了扩散模型的效率。

见证了扩散模型在图像生成领域的成功，许多方法也开始将扩散模型的框架迁移到其他的生成任务上，如视频生成[26]、音频生成[101]，甚至文本生成[102]。在迁移的过程中，往往需要根据目标任务的特点进行适应性的重构[103, 104]。例如，为了解决预测结果为离散标签的任务，Chen 等人[103] 提出将二进制位将离散任务标签转换为连续状态，从而更好地适配连续的扩散过程。同时，也有一部分方法[105] 考虑把扩散模型的框架迁移到多任务领域。具体来说，它将一些任务看作辅助任务作为额外输出，并利用其引导图像生成过程，而生成过程中产生的内部特征会被用来预测任务类型。当期在处理不同类型的任务时，它使用了不同类型的编码器将不同任务标签编码到扩散模型特征空间。这类方法尽管把扩散模型向多任务的领域进行了扩展，但是其基本思路还是基于主任务和辅助任务，因此只能同时处理两个任务，在任务处理的数量上还需要进一步的改进。

在另一方面，不止于生成式任务，扩散模型强大的建模能力也让研究者们开始探索其在感知类任务上的应用。尤其在密集预测任务方面，许多方法[23, 27, 43–45] 都在尝试将扩散模型应用于不同的密集预测任务中，并且都取得了突出的效果。举例而言，Wang 等人[23] 把扩散模型用于建立通用分割模型，把图像分割看成是给定图像作为条件后对于特定类别掩码的生成任务，并从高斯噪声中一步步去噪从而得到最终的结果。Saxena 等人[44] 把扩散模型用于估计图像深度，把单目深度估计看作是一种特殊的图像进行生成。近年来，诸如 Stable Diffusion 等方法[106] 利用大量的模型参数和训练数据，以潜空间扩散模型为基础，构建具有强大图像生成能力的模型，因此也有方法[107, 108] 利用其预训练的权重构建密集预测方法。具体来说，Ke 等人[107] 对 Stable Diffusion 进行微调，将其输入在带噪声的潜空间特征的基础上增加了作为条件的图像潜空间特

征（这一特征由变分自编码器[109] 编码而来），并且将预测目标由图像修改为深度预测图。这一方法仅仅使用少量的训练数据就能够表现出强大的零样本效果。类似的，Zhu[108] 等人用同样的方法将 Stable Diffusion 的预训练权重用于微调少样本语义分割任务，同样取得了突出的效果。尽管这些方法适用的领域有所不同，但是它们都有着类似的流程，即首先使用编码器编码图像特征，然后将这一特征作为条件输入解码器中，辅助解码器对带噪声的预测结果进行去噪。基于此，Ji 等人[27] 将多个密集预测任务统一到了一个简洁的扩散模型框架中，利用编码器获取条件特征，并用解码器对带噪声的预测结果进行解码。为了提升模型的效率，该方法将解码器轻量化，因此迭代的去噪过程也不会对模型的推理产生过高的计算负担，提升了模型的效率。但是，上述的这些方法专注于一个模型解决单个任务，在多任务密集预测领域的相关研究目前相对较少。现有的多任务密集预测方法里，Ye 等人[110] 利用扩散过程校正部分标注的多任务学习中含有大量噪声的预测结果。该方法仅聚焦于处理因部分监督标签导致的噪声预测去噪问题，却忽视了在全标注场景下扩散模型解决多任务密集预测问题本身的潜力。同时，这一方法忽略了多任务密集预测上扩散模型的效率问题。与这一方法不同，本文所提出方法致力于用扩散模型建模任务与任务之间的关系，同时设计高效的去噪过程，从而缓解多任务场景下迭代去噪的效率问题。

第三章　基于低秩混合专家的多任务密集预测

第一节　背景

计算机视觉任务，如语义分割[2–5] 和深度估计[7, 8], 在深度学习技术推动下已取得显著进展。基于深度学习的计算机视觉方法通常采用精心设计的深度模型，尽管不同任务的模型结构可能有所不同，但通常遵循相似的流程，包括特征提取和根据特征进行预测。此外，一些任务之间也存在关联，而这样的关联能够在不同任务的特征之间形成互补，相互提升各自的性能。

这些事实促使研究人员研究多任务学习，也就是够将不同的任务模型结合到一个模型中的技术。相较于单任务模型，多任务学习的显著优势在于，它能在保持各任务模型性能的同时，提升训练和推理效率并减少参数负担。由于这一优势，多任务学习模型已被应用于多个方向，帮助前沿视觉模型在边缘设备上进行部署，如自动驾驶[9–11] 和场景理解[18, 19] 等。

本章聚焦于密集场景理解中的多任务学习，并且从模型架构这一方面入手对现有方法的局限与本章所提出的改进进行进一步讲解。在多任务密集预测里，早期工作的一条研究路线[13, 14, 17–19, 30, 78] 侧重于设计精细的网络架构，这一路线可以被更具体地区分为以编码器为主的方法和以解码器为主的方法。以编码器为主的方法[13, 14, 78] 设计手工制作的模块以在特定任务的编码器之间共享，构建任务通用特征，而解码器为主的方法[17–19, 30] 则偏向于定制解码器，以学习更具有区分度的任务专属特征，并构建跨任务关系。

与上述专注于设计静态网络架构的方法不同，一些方法[33–35] 引入了混合专家模型（MoE）技术，提供了一种动态自动学习方式来学习参数之于任务的专属或共享[34]。具体来说，它们利用 MoE 设计编码器模块，并为不同任务和样本动态选择网络路径。相较于这些编码器为主的方法，MoE 在解码器中的应用研究相对较少。

最近，Ye 等人[15] 首次将 MoE 应用于解码器，该方法通过动态组合来自不同专家的任务通用特征解码任务专用特征，从而提升了任务专用特征之间的区分度和多样性，并且在性能上优于以前的解码器为主的方法。这一成功推动了

表 3.1 标准 MoE 和低秩混合专家模型在不同设置下的参数和 FLOPs。左侧设置展示了专家数量和专家网络中卷积核大小。

设置	参数 (M)		FLOPs (G)	
	MoE	MLoRE	MoE	MLoRE
5 个专家, [1×1, 1×1]	3.1	1.2	3.00	1.49
10 个专家, [1×1, 1×1]	4.7	1.6	4.49	1.58
15 个专家, [1×1, 1×1]	6.3	1.9	5.99	1.66
5 个专家, [3×3, 1×1]	14.9	3.4	14.24	7.12
10 个专家, [3×3, 1×1]	22.4	4.7	21.37	7.21
15 个专家, [3×3, 1×1]	29.9	6.0	28.49	7.29

本领域的研究者们对基于 MoE 的多任务学习解码器的深入研究。

得益于动态路由过程，这些基于 MoE 的方法能显著提高参数和特征的多样性，从而产生更具区分度的任务专属特征。然而，该模式仍存在若干局限。首先，虽然基于 MoE 的方法能通过在动态路由过程中共享相同专家来在部分任务中建立连接，但在所有任务之间共享专家的机会较低，这可能会阻碍路由器在所有任务之间建立全局关系。同时，全局关系建模在之前的方法[19, 81] 中已被证明对密集型多任务学习有重要作用。由此可见，在 MoE 中明确建模所有任务之间的全局关系是至关重要的。此外，任务通用特征空间的容量与专家数量密切相关。在现有方法[88, 111] 的实验中，可以看到增加专家数量有助于提升模型的容量，从而促进不同任务的整体性能。然而，专家数量的增加会导致参数量和计算开销的显著提升，这对现有多任务密集预测方法构成沉重负担。

针对上述问题，本章提出了一种新的解码器为主的方法，称之为低秩混合专家模型（MLoRE）。MLoRE 框架的核心思想是显式建模 MoE 中所有任务之间的全局关系，并在增加专家数量以扩大模型容量时减轻 MoE 的计算负担。针对全局关系建模问题，MLoRE 在标准 MoE 结构基础上引入了与 MoE 模块并行的任务共享卷积路径。

具体来说，MLoRE 首先将骨干网络特征映射为不同的任务专属特征，然后将它们全部输入到通用卷积路径和原始 MoE 的专家网络中。通过在所有任务中共享相同的通用路径，MLoRE 得以建立起所有任务之间的关系。此外，为了增强任务专用特征的区分性，MLoRE 还设置了专用于特定任务的专家网络，这些专家不参与动态路由机制。

在效率优化方面，本章从低秩适应中获得灵感，认为适应不同任务的基本

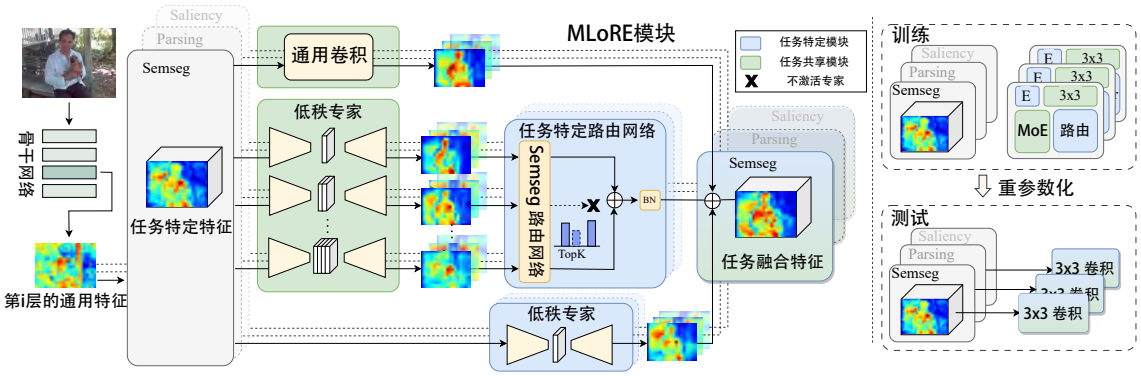

图 3.1 本章方法的整体框架。MLoRE 模块被装配在不同的层级上，来自不同层级的主干特征分别输入到 MLoRE 模块中。在每个选定的层级上，首先将主干特征映射到不同的任务专属特征中，然后送入任务共享卷积、任务共享低秩专家网络，接着经过任务特定的路由网络和任务特定的低秩专家网络。这些分支的输出被累加以生成对应任务的任务专属特征。在每个选定的层级上堆叠了两个 MLoRE 模块。

模型仅需要低秩权重更新。在此基础上，本章将 MoE 的专家网络转换为普通卷积权重低秩分解之后的形式，其效果如表 表 3.1所示，相比标准混合专家模块减少了 60% 以上的参数消耗。

此外，为了控制由于增加专家网络数量带来的计算成本，MLoRE 通过移除所有非线性激活函数来支持推理阶段的重参数化。经由重参数化，不同专家网络的知识被注入到通用卷积路径中，从而减少密集预测任务的计算成本。值得强调的时，MLoRE 是首个在 MoE 中使用线性专家进行多任务密集预测的工作。为了验证 MLoRE 的有效性，本章在 PASCAL-Context 和 NYUD-v2 数据集上进行了广泛实验。

总结起来，本章贡献主要有三点：

- 本章分析了 MoE 在多任务学习中应用时遇到的问题，并提出了一种全新的以解码器为中心的框架 MLoRE，能够显式建模所有任务之间的全局关系，并在不显著增加模型规模的情况下扩展特征表示的能力。
- 本章引入了一个简单的任务共享通用路径到 MoE 结构中，并提出了基于低秩适应启发的线性和低秩专家网络。同时，通用卷积路径和低秩专家路径可以通过线性结合在推理时进行重参数化，进一步保障效率。
- 在 PASCAL-Context 和 NYUDv2 上的实验表明，本章所提出的方法在所有任务上明显优于之前的最先进的多任务学习方法。

第二节 方法

一、 整体框架

整体框架遵循了之前工作中使用的多尺度架构[15, 81]。具体来说，本章使用视觉 Transformer（ViT）作为编码器，并从不同层级收集多尺度特征。在公式化表达下，给定输入图像 $\mathbf{I}$ 和一个视觉 Transformer $\mathcal{F}$, 可以从不同层级中获得多尺度特征集 $\left\{\mathbf{X}^l = \mathcal{F}^l(\mathbf{I})\right\}$，其中 $\mathbf{X}^l \in \mathbb{R}^{L\times C}$。这里，$l$ 表示层级索引, $L = H \times W$ 表示图像块的数量, C 表示特征维度。$\mathcal{F}^l(\mathbf{I})$ 是第 l 层 Transformer 的输出特征。提取出来的多尺度特征会被输入到解码器中，该解码器包括每个尺度上堆叠的两个 MLoRE 模块。对于每个任务，从不同尺度的 MLoRE 模块的输出特征会被拼接在一起，以生成最终的密集预测任务特征。

二、 低秩混合专家模型

在具体描述所提出的低秩混合专家（MLoRE）模块之前，本章首先介绍 MoE 的基本形式 $f_{moe}(\cdot)$。形式上，假设 MoE 包含 N 个专家和 T 个路由网络, 分别记作 $\mathbb{E} = \{E_1，E_2，...，E_N\}$ 和 $\mathbb{G} = \{G_1，G_2，...，G_T\}$ 。N 和 T 分别是专家的数量和任务的数量。来自第 l 层的骨干网络特征 $\mathbf{X}^l$ 被分别输入到 N 个专家网络和 T 个路由网络中。为了方便起见，下面的公式中省略了上标 l。对于第 n 个专家，判别输出特征由 $\mathbf{X}n = E_n(\mathbf{X})$ 生成。与此同时，MoE 从任务特定的路由网络中学习门控值用于不同的任务。对于任务 t，由路由网络生成的每个专家的门控值可以表示为 $\mathbf{g}^t = G_t(\mathbf{X})$，其中 $\mathbf{g}^t \in \mathbb{R}^N$。最后，MoE 利用门控值结合专家特征来生成任务 t 的任务专属特征 $\mathbf{S}^t$，其公式为

$$\mathbf{S}^t = f_{moe}(\mathbf{X}) = \sum_{n=1}^{N} \mathbf{g}_n^t \mathbf{X}_n。 \tag{3.1}$$

对于 M^{3}ViT[33] 和 Mod-Squad[34]，它们根据相应的门控值在一次推理中关闭一些专家，并选择前 k 大专家。这些任务专属的特征用于对每个任务进行预测。

MoE 的优点在于可为每个样本和每个任务动态地编码特征，并通过多个专家增加特征编码的多样性。然而，研究发现将 MoE 技术应用于构建多任务学习解码器时，其难以有效建模全局任务关系。此外，当增加专家数量扩大特征表示能力和专家网络的上下文时，参数和计算成本也会相应增加。针对这些问题，本章提出了低秩混合专家（MLoRE）模块。

本章所提出的低秩混合专家（MLoRE）模块的整体流程如 图 3.1 所示。为

了在任务专属特征之间构建跨任务关系，本章首先使用几个轻量级的卷积层将骨干网络特征投射为不同任务的任务专属特征。然后，每个任务的特征被送到任务共享的通用路径和多个任务共享的专家网络，专家网络由具有显式低秩约束的卷积层构成。对样本进行处理时，低秩专家的选择是根据任务专属的路由网络对每个任务的预测结果来决定的。此外，除了基于任务专属路由器构建任务特定特征外，本章还引入了额外的任务专属低秩专家网络，以帮助构建更具区分度的任务专属特征。来自任务共享通用路径、任务共享低秩专家网络（由任务特定路由器选择）和任务专属专家网络的特征被加总在一起，以生成具有区分度的任务特定特征。最后，本章在 MLoRE 模块中去除了所有非线性模块，即在所有路径中都不使用任何激活函数，从而实现重参数化以减少计算成本。

具体来说，第 l 层的骨干网络特征 $\mathbf{X}$ 首先通过 1×1 卷积映射到每个任务对应的任务专属特征中，这可以表示为 $\left\{\mathbf{X}^t = f_{t,1\times1}(\mathbf{X})，t\in[1，\cdots，T]\right\}$，其中 $\mathbf{X}^t\in\mathbb{R}^{C\times H\times W}$。然后，各个任务的专属特征分别被送到三个路径中，即任务共享的通用路径 $f_g(\cdot)$、任务共享低秩专家路径 $f_{lre}(\cdot)$（具有任务专属路由网络 $f^t_{sr}(\cdot)$）和任务专属低秩专家路径 $f_{se}(\cdot)$。最终的任务专属特征 $\mathbf{S}^t$ 通过以下公式获得

$$\mathbf{S}^t = f_g(\mathbf{X}^t) + f^t_{sr}\left(f_{lre}(\mathbf{X}^t)\right) + f^t_{se}(\mathbf{X}^t)。\tag{3.2}$$

在本章的方法中，每个选定的骨干网络层后会堆叠两个 MLoRE 模块。在第一个 MLoRE 模块中，轻量级的任务特定 1×1 卷积被用来将骨干特征投影到不同的任务专属特征。在第二个 MLoRE 模块中，由于任务专属特征已经被区分开，本章直接使用 1×1 卷积来处理任务专属特征。此外，由于 MLoRE 是一个线性模块，所以本章在两个 MLoRE 模块之间添加了一个任务专属的非线性块，以将非线性引入解码器。每个非线性块由“GELU-BatchNorm-线性结构”组成。

接下来, 本章介绍 MLoRE 模型中关于这三个路径的网络细节.

任务共享通用路径：任务共享通用路径包含一个 3×3 卷积层，其权重矩阵为 $\mathbf{W}_g\in\mathbb{R}^{3\times3\times C\times C}$ 和偏置矩阵 $\mathbf{b}_g\in\mathbb{R}^C$。由于所有任务特征都会经过此通用卷积，因此它会同时通过不同任务的梯度进行优化，这有助于提取所有任务之间的共享特征。为了让模型的优化更加顺利，本章在训练过程中会停止对该路径的梯度进行进一步反向传播。因此，梯度只会通过其他两个路径进行反向传播。本章发现这样一个简单的操作可以更好地缓解优化过程中的梯度冲突。如在 第三节 中的实验结果所示, 本章所提出的任务共享通用路径可以在所有任务上带

来性能提升，证明了从全局角度明确建立跨任务关系的想法的有效性。

任务共享低秩专家路径：本章借鉴了低秩适应[96]的思想，采用低秩卷积，也就是在普通卷积上显式地增加低秩约束。具体来说，每个任务共享的专家网络包括一个 3×3 卷积和一个 1×1 卷积，两个卷积的权重可以被重参数化为一个低秩的卷积权重矩阵，每个任务共享专家网络具有类似的结构。所有任务共享专家网络的权重和偏置可以表示为 $\left\{\mathbf{W}_{lreb}^{n}，\ \mathbf{b}_{lreb}^{n}，\ \mathbf{W}_{lrea}^{n}，\ \mathbf{b}_{lrea}^{n} | n \in [1，\ ...，\ N]\right\}$，其中 $\mathbf{W}_{lreb}^{n} \in \mathbb{R}^{3\times3\times C\times r_n}$，$\mathbf{b}_{lreb}^{n} \in \mathbb{R}^{r_n}$，$\mathbf{W}_{lrea} \in \mathbb{R}^{1\times1\times r_n\times C}$，且 $\mathbf{b}_{lrea}^{n} \in \mathbb{R}^{C}$ $(r_n \ll C)$。r_n 表示第 n 个专家网络的秩。在本章的方法中，不同专家网络的 r_n 值不同，这一设计旨在提高参数和特征的多样性。对于每个任务，任务特定路由网络 $f^t sr(\cdot)$ 学习这些专家的门控值并根据门控值激活前 k 大专家。所有被激活的专家的输出特征会进行加和，加和的结果送入 BatchNorm 层生成任务特定特征。BatchNorm 层包含四个参数，包括累积的通道均值 $\bar{} \in \mathbb{R}^{C}$、累积的通道标准差 $\mathbf{œ} \in \mathbb{R}^{C}$、缩放因子 $\gamma \in \mathbb{R}^{C}$ 和偏置 $\beta \in \mathbb{R}^{C}$。

任务专属低秩专家路径：包含 T 个任务专属的专家网络，每个网络负责其对应的任务特征。对于每个任务专属专家网络，本章使用与任务共享专家路径类似的结构，包括一个 3×3 卷积，其权重矩阵为 $\mathbf{W}_{seb}^{t} \in \mathbb{R}^{3\times3\times C\times R}$ 和偏置矩阵 $\mathbf{b}_{seb}^{t} \in \mathbb{R}^{R}$，接着是一个 1×1 卷积，其权重矩阵为 $\mathbf{W}_{sea}^{t} \in \mathbb{R}^{1\times1\times R\times C}$ 和偏置矩阵 $\mathbf{b}_{sea}^{t} \in \mathbb{R}^{C}$。$R$ 表示秩数 $(R \ll C)$。任务特定专家路径可以增强任务特定特征的区分度，这一点也将在后续的实验中验证。

路由网络：如 图 3.1 所示，为了从任务共享低秩专家路径生成任务专属特征，本章设置了任务专属路由网络，另其为每个专家生成门控值，并将其作为不同专家特征输出的线性组合的权重。每个任务的路由网络通常是简单的线性层，后跟平均池化层和预测层。具体来说，任务 t 的路由网络 $f_{sr}^{t}(\cdot)$ 设计如下。本章的路由网络以任务专属特征 $\mathbf{X}^{t} \in \mathbb{R}^{C\times H\times W}$ 作为输入，并将其送入两个连续的 1×1 卷积，将通道维度从 C 映射到 $C/4$，然后是一个全局池化层。最终的输出是一个全局特征向量 $\mathbf{X}_f \in \mathbb{R}^{\frac{C}{4}}$。

此外，受之前工作[112, 113]的启发，本章引入了另一个并行的基于位置的分支。类似地，它由两个线性层组成。第一个线性层沿空间维度缩小特征，将形状从 $\mathbb{R}^{C\times HW}$ 映射到 $\mathbb{R}^{C\times1}$，然后通过第二个线性层转换到 $\mathbb{R}^{\frac{C}{4}}$。这两个分支的输出特征向量沿通道维度拼接在一起，然后送入最终预测层，并通过 Softmax 函数以生成每个专家的门控值 g_t 。

推理时的重参数化：本章通过移除所有激活函数，在 MLoRE 模块中引入了线性化，这使得在推理时可以将所有路径的参数重参数化为每个任务的简单 3×3 卷积。本章首先对任务共享低秩专家路径进行重参数化，然后对所有路径的参数进行重参数化。根据[114] 的研究，任务共享专家路径中的权重和偏置矩阵可以合并并表示为：

$$\mathbf{W}_{lre}^{t} = \mathfrak{B}(\frac{\gamma}{\sigma}) \sum_{k \in \mathbb{K}_t} \mathbf{g}_{k}^{t} \mathbf{W}_{lreb}^{k} \mathbf{W}_{lrea}^{k}, \tag{3.3}$$

$$\mathbf{b}_{lre}^{t} = \frac{\gamma}{\sigma}(\sum_{k \in \mathbb{K}_t} \mathbf{g}_{k}^{t} (\mathbf{b}_{lreb}^{k} \mathbf{W}_{lrea}^{k} + \mathbf{b}_{lrea}^{k}) - \mu) + \beta, \tag{3.4}$$

其中，$\mathfrak{B}$ 表示广播操作，$\mathbb{K}_t$ 表示由路由网络为任务 t 选择的激活专家的索引集。$\mathbf{g}_{k}^{t}$ 是路由网络预测的第 k 个门控值。在推理阶段，这三个路径的权重矩阵和偏置矩阵可以重参数化为：

$$\mathbf{W}_{r}^{t} = \mathbf{W}_{g} + \mathbf{W}_{sr}^{t} + \mathbf{W}_{seb}^{t} \mathbf{W}_{sea}^{t}, \tag{3.5}$$

$$\mathbf{b}_{r}^{t} = \mathbf{b}_{g} + \mathbf{b}_{sr}^{t} + \mathbf{b}_{seb}^{t} \mathbf{W}_{sea}^{t} + \mathbf{b}_{sea}^{t}。 \tag{3.6}$$

$\mathbf{W}_{r}^{t}$ 和 $\mathbf{b}_{r}^{t}$ 是重参数化卷积的权重和偏置。因此，公式 (3.2) 可以重新表述为：

$$\mathbf{S}^{t} = \mathbf{X}^{t} \circledast \mathbf{W}_{r}^{t} + \mathfrak{B}(\mathbf{b}_{r}^{t}), \tag{3.7}$$

其中，$\circledast$ 表示卷积操作，$\mathbf{b}_{r}^{t}$ 通过广播机制与 $\mathbf{X}^{t}$ 具有相同的形状。

此外，由于重参数化可以加速前向传播，因此是否可以将重参数化扩展到训练阶段以提高训练效率就是一个非常自然的问题。然而由于网络设计的原因，在本章的 MLoRE 模块中，重参数化只能在推理阶段进行。具体来说，在训练阶段的重参数化会极大地影响训练时的网络行为，而其原因在于 MLoRE 模块中的 BatchNorm 层。本章遵循 RepVGG[115] 的方法，在任务共享的低秩专家支路中设置 BatchNorm。而根据 RepVGG 中的实验结果表明，BatchNorm 层对于基于重参数化的方法来说非常重要，所以无法移除该层。同时，当 BatchNorm 层在训练中与卷积层合并时，该 BatchNorm 层的特征统计将难以执行。综上所述，本章的重参数化过程只能在推理中进行。

MoE 的优化方式遵循之前基于 MoE 的多任务学习（MTL）方法[33, 34]，本章采用了 Shazeer 等人[116] 提出的噪声门控和负载均衡损失，这是稀疏 MoE 训练中的常见做法[116, 117]。

在没有负载均衡损失的情况下，会存在同一个样本上的所有任务更激活同一个专家的可能，而这正好与任务共享支路的作用重合。但是，移除负载均衡损失来在一个专家中构建所有任务的全局关系却并非一个好的设计，其原因有二。首先，一处负载均衡损失会削弱 MoE 为不同样本动态选择不同专家的能力，这与本章使用 MoE 的初衷相悖。其次，如果没有负载均衡损失，大多数专家将很少或从未被激活，这同样会损害 MoE 的能力。而与上述的情况相反，本章提出的任务共享通用路径不会损害动态路由能力和 MoE 的容量。本章将在后续实验中证明负载均衡损失的必要性。

此外，本章的 MLoRE 在训练时采用了前 k 大约束。在没有前 k 大约束的情况下训练 MoE 时，本章发现每个专家会被所有任务共享，从而构建全局任务关系。然而通过实验可以发现，这可能会使优化过程变得十分困难，并损害 MoE 在任务子集中构建关系的能力。因此，尽管它可以构建全局任务关系，但如果没有前 k 大约束，性能会受到很大影响，这一点会在后续的实验中得到体现。与此相对的，本章提出的任务共享通用路径可以显式地构建全局任务关系，而 MoE 仍然可以构建任务子集之间的关系。

第三节 实验

一、 实验设置

数据集：为了证明本章方法的有效性，本章在两个流行的多任务数据集上评估了 MLoRE 的性能，这两个数据集是 PASCAL-Context[118] 和 NYUD-v2[119]。PASCAL-Context[118] 包含了多个任务的高质量标注，包括语义分割、人体解析、显著性检测、表面法线和物体边界检测。该数据集中有 4,998 张训练图像和 5,105 张测试图像。NYUDv2[119] 也提供了高质量的多任务注释，包括语义分割、单目深度估计、表面法线和物体边界检测。该数据集包含 795 张训练图像和 654 张测试图像。

评估指标：下面，本章将介绍上述任务的评估指标。遵循以往的多任务工作[19, 81]，本章使用平均交并比（mIoU）来评估语义分割和人像解析的性能，使用根均方误差（RMSE）用于评估单目深度估计的准确性，显著性检测使用最大 F-measure（maxF）来评估，表面法线和物体边界检测分别采用平均误差（mErr）和最优数据集尺度 F-measure（odsF）作为评估指标。为了评估所有任务的总体性能，本章按照[120] 的方法评估了所有任务的 MTL 增益 Δ_m。在后续实验结果

表 3.2 在 PASCAL-Context 数据集上 MLoRE 不同组件的消融研究。每一行在上一行的基础上增加了一个额外的设置。MoE：标准的混合专家结构；LoRE：任务共享低秩专家路径；GC：任务共享通用卷积路径；SPE：任务特定专家路径。↑表示越高越好。↓表示越低越好。

设置	Semseg	Parsing	Sal.	Normal	Bound.	MTL	FLOPs	# 参数量
	mIoU ↑	mIoU ↑	maxF ↑	mErr ↓	odsF ↑	Δ_m ↑	(G)	(M)
基线方法	77.38	65.15	85.08	13.79	69.87	-3.41	391	115
+ MoE	78.56	66.78	85.18	13.57	73.91	-1.20	1834	676
基线方法								
+ LoRE	78.38	66.21	85.15	13.71	73.53	-1.71	568	213
+ GC	79.25	67.43	85.20	13.70	74.38	-0.88	568	243
+ SPE	79.26	67.82	85.31	13.65	74.69	-0.58	568	259

中，本章用 semseg 代表语义分割任务，parsing 代表人体解析任务，saliency 或者 sal. 代表显著性物体检测任务，normals 代表法线检测任务，edge 或者 Bound. 代表边缘检测任务，depth 代表深度估计任务。

训练设置：本章使用 ViT-large[73] 作为骨干网络，同时将解码器的通道数设置为 384。在消融研究力，骨干网络被设置为 ViT-base 网络。按照之前的工作[81]，本章在这两个数据集上把批大小设置为 4，并训练了 40,000 次迭代。不同任务的优化器和损失函数遵循之前的工作[81] 中的设置。

二、 消融实验

在本小节中，本章进行广泛的实验来展示不同组件的有效性，并找到不同超参数的最佳设置。除非另有说明，本小节所有的消融实验都是基于 ViT 骨干网络进行的。本方法基线建立在具有 12 层的 ViT-base 主干网络上，利用来自第 3 层、第 6 层、第 9 层和第 12 层的骨干网络特征作为多尺度特征，每个特征后面都跟着一个线性层，将通道维度投影到每个任务的输出通道。

不同组件的有效性：本章首先进行实验，以验证 MLoRE 模块中不同组件的有效性。表 3.2显示了定量结果。本章首先检查带有标准 MoE 的基线网络的性能，以及它们模型的参数大小和 FLOPs。标准 MoE（15 个专家网络）中的专家网络与本章的相似，每个网络都由一个 3×3 卷积和一个 1×1 卷积组成，中间使用 ReLU 激活函数，但是并没有显式的低秩约束。当将 MoE 添加到 Baseline 中时，本章观察到 MTL 增益有所提高，但参数和 FLOPs 也分别增加了大约 5 倍和 4 倍，这对整个网络来说是一个沉重的负担。当将低秩专家网络（LoRE）添加到基线网络时，性能同样得到了提升，但参数和 FLOPs 仅为基于 MoE 模型的

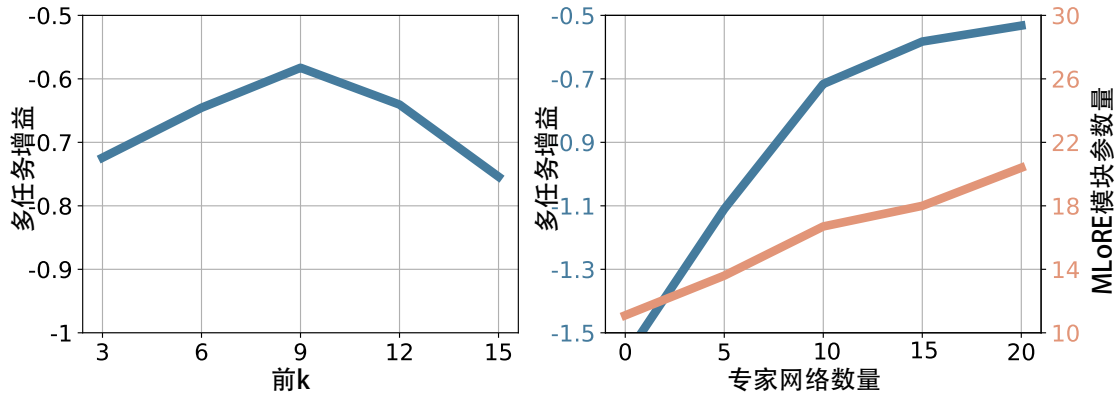

图 3.2 关于专家数量 N 和激活专家数量 K 的消融实验。右图还展示了随着专家数量增加，MLoRE 模块参数量的变化的趋势。

1/3。当将低秩特性应用于专家网络时，参数大小减少了数倍。同时，通过移除所有激活函数，在专家网络中引入线性化，可以通过将所有专家重参数化为单个卷积来节省计算资源，其有效性同样在表中得到了体现。

此外，本章还强调在 MoE 中明确构建全局任务关联的重要性，并引入任务共享通用路径来实现这一目标。可以看到，向 LoRE 添加任务共享通用路径可以进一步提升性能，并在大多数指标上超越带有 MoE 的基线网络，这证明了建模所有任务之间全局关系的有效性。此外，为每个任务添加任务特定的低秩专家也提升了性能，证明了任务专属的专家网络可以增强任务专属特征的区分性。本章经验性地将任务专属低秩专家的秩设为 64。

任务共享低秩专家的数量和 top-k 选择：本章对 MLoRE 模块中低秩专家的数量和任务专属路由网络选择的前 k 大专家进行了消融实验。本章首先固定一个参数，并消融另一个参数，以研究它们对多任务性能的影响。如图 3.2所示，当增加专家数量时，模型的 MTL 增益显著提升，并在专家数量为 15 时实现了最佳性能。当进一步增加专家数量时，无法观察到明显的性能提升。此外，当将专家数量固定为 15 时，本章对激活专家的比例进行了消融实验。可以观察到，在这一系列实验中，为每个任务激活 60% 的专家是最佳选择。当选择所有专家网络时，性能大幅下降，这反映了稀疏性对于特征区分的重要性，也应证了前文中需要任务共享支路来建立所有任务间关系的必要性

秩数设置：专家网络利用的是带有 640 个输出通道的标准 3×3 卷积的低秩格式。在专家网络中，不同专家网络权重的秩 r 也起着重要作用。本章研究了不同的设置，包括 1) 所有专家的秩数为 16，2) 所有专家的秩数为 128，3) 专家秩

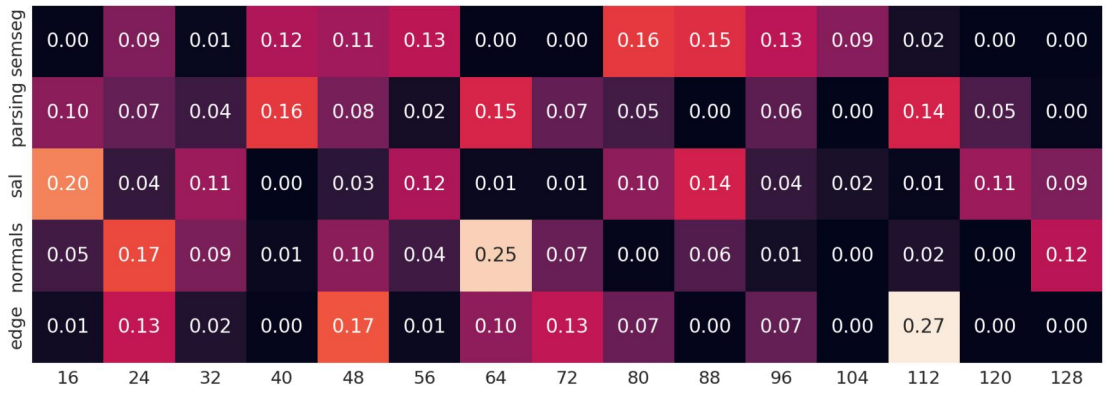

图 3.3 任务与低秩专家之间的关系图。

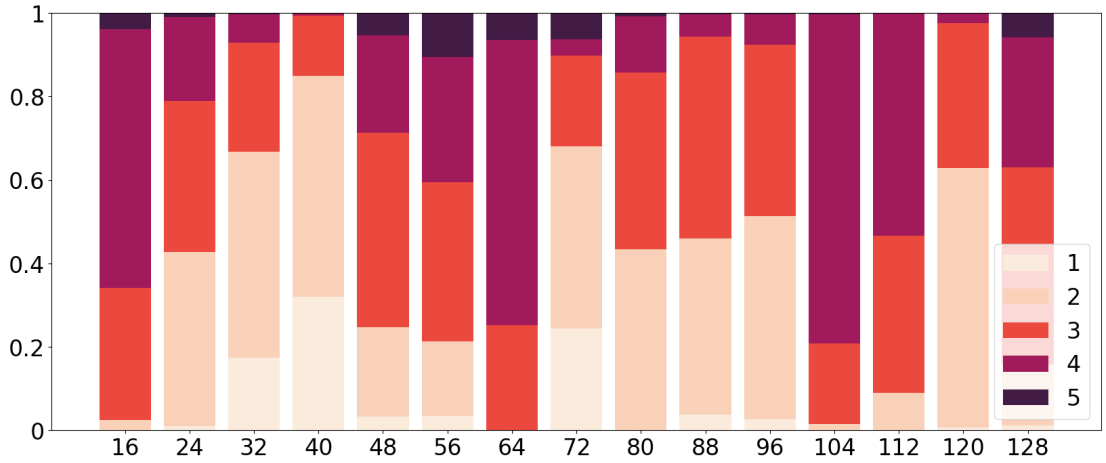

图 3.4 MLoRE 模块中不同任务激活专家的比例（未使用任务共享通用路径）。可以看到，在没有任务共享通用路径的情况下，只有少数专家能够被所有五个任务激活。横坐标表示不同专家的秩。

数从 16 增加到 128，每次增加 8 个单位。如表 3.3所示, 可以看到为专家网络选择不同的秩数可以实现最佳的 MTL 增益，可以带来比相同秩数更多的特征多样性，因此本章将其作为方法最终的设置。**关系可视化：**任务与低秩专家之间的关系如 图 3.3所示。本章统计了最后阶段第二个 MLoRE 模块中的关系，并计算了不同任务在整个数据集上选择的每个专家的激活比例。可以看到，不同秩数的专家倾向于学习不同的任务子集。具体而言，秩较低的专家倾向于为 3-4 个相关任务学习共享知识，而秩较高的专家则专注于 1-2 个任务。此外，本章还展示了在未添加任务共享通用路径时，MLoRE 模块中不同任务激活不同专家的比例，如图 3.4所示。可以看到，在完全动态的方式下，这些专家很少或从未被所有任务在同一个样本中激活。这证明了在直接使用 MoE 解码器时，几乎没有专

表 3.3 在 PASCAL-Context 数据集上，MLoRE 任务共享低秩专家路径中的秩设置消融实验。

最小/最大秩	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Δ_m ↑
16/16	78.84	68.01	85.32	13.69	74.39	-0.77
16/128	79.26	67.82	85.30	13.65	74.69	-0.58
128/128	78.79	66.98	85.35	13.67	74.29	-1.07

表 3.4 路由网络设置的消融实验。**basic**: 基础路由网络。**pos.**: 位置感知路由网络。**w/o sample-dep**: 输入为样本无关的可学习参数。

路由网络种类	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Δ_m ↑
basic	79.15	67.40	85.21	13.58	74.34	-0.75
basic+pos.	79.26	67.82	85.31	13.65	74.69	-0.58
only pos.	79.10	67.76	85.11	13.71	74.51	-0.82
basic	79.15	67.40	85.21	13.58	74.34	-0.75
w/o sample-dep	78.86	67.38	85.41	13.65	74.25	-0.91

家可以学习所有任务的全局关系。这种现象有力地支持了任务共享通用路径和显式建模全局任务关系的必要性。

低秩任务共享通用路径: 本章进一步验证了低秩任务共享通用路径的有效性。考虑到本章的方法在探索低秩约束在 MoE 中的有效性，本章考虑在任务共享通用路径中对 3×3 卷积进行显式的低秩约束，以探索是否可以在 MLoRE 中设计一个完全低秩结构的更轻量模块。结果如表 3.5所示。随着秩的增加，大多数任务的性能都有所提升。当使用普通 3×3 卷积时，其性能显著优于低秩设置。这一结果表明，使用没有低秩约束的 3×3 卷积来构建任务共享通用路径在性能上有着正面的影响，因此也支持本章在实际的模型中使用正常卷积层而非其具有低秩约束的形式。

任务特定的路由网络： 路由网络对于生成任务特定的门控机制至关重要，这决定了如何激活专家并组合他们的特征。本章对路由网络的几个模型设计上的选项进行了消融实验，结果如 3.4 所示。在基础路由网络中添加位置感知分支可以将 MTL 增益提高 +0.17。这表明位置感知分支可以获得更多的上下文信息，对路由网络有益。此外，当将路由的输入从可学习参数更换为样本特征时，MTL 增益增加了 +0.16, 这表明样本的动态信息对门控机制至关重要。

负载均衡损失的有效性： 本章进行了大量实验以验证负载均衡损失的有效

表 3.5 在 Pascal-Context 进行的关于低秩任务共享支路的消融实验。

秩	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Δ_m ↑
16	78.53	66.71	85.27	13.68	73.70	-1.41
128	78.78	67.01	85.08	13.70	73.98	-1.26
满秩 (默认设置)	79.26	67.82	85.30	13.65	74.69	-0.58

表 3.6 在 Pascal-Context 数据集上不同设置下对于负载均衡损失的消融实验。**MoE**: 使用基础 MoE 结构的基线网络。**LoRE**: 在基础 MoE 结构基础上增加任务共享支路的基线网络。**Ours**: 带有所有组件的 MLoRE 网络。**w/o LB loss**: 移除负载均衡损失。

设置	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Δ_m ↑
MoE	78.56	66.78	85.17	13.58	73.91	-1.20
w/o LB loss	78.32	66.41	85.14	13.58	73.81	-1.40
LoRE	78.38	66.21	85.15	13.71	73.53	-1.71
w/o LB loss	78.04	66.05	85.10	13.69	73.60	-1.80
Ours	79.26	67.82	85.30	13.65	74.69	-0.58
w/o LB loss	79.01	68.03	85.24	13.66	74.38	-0.70

性。为了证明其对 MoE 结构的必要性，本章测试了三种不同设计的设置。结果如表 3.6所示。可以清楚地看到，在使用负载均衡损失的情况下，所有三种设置在大多数任务上都能取得更好的性能。负载均衡损失还提高了多任务学习（MTL）的增益。实验的定量结果证明了负载均衡损失对 MoE 结构的有效性，并启发本章提出任务共享通用路径，而不是移除负载均衡损失来构建全局关系。

三、 与其他方法的比较

与之前最先进的（SOTA）方法的定量比较如表 3.7 和 表 3.8所示。可以看出，本章的方法在 PASCAL-Context 和 NYUDv2 数据集上的所有指标上明显优于之前的方法。尤其是在 PASCAL-Context 数据集上，语义分割、人体解析和边界检测的表现分别比之前最好的方法提升了 +0.52 mIoU、+1.10 mIoU 和 +1.92 odsF。为了更加直观地展现本章方法的性能，本章将不同方法的性能在 图 3.5中可视化，可以看到本章方法在所有任务上都有十分显著的提升。

之前的方法，如 M^3ViT[33]、Mod-Squad[34] 和 TaskExpert[15] 都在他们的网络中使用了 MoE 技术。然而，本章的方法表现优于它们，这证明了 MLoRE 模块的有效性。相比于专注于解码器的方法 TaskExpert，本章在语义分割、人体解

表 3.7 PASCAL-Context 数据集上不同方法的定量比较。* 表示基于[15] 中 ViT-large 骨干网络复现的方法性能。

方法	骨干网络	**Semseg** mIoU ↑	**Parsing** mIoU ↑	**Saliency** maxF ↑	**Normal** mErr ↓	**Boundary** odsF ↑	**FLOPs** (G)	**# 参数量** (M)
PAD-Net[18]	HRNet18	53.60	59.60	65.80	15.30	72.50	124	81
MTI-Net[17]	HRNet18	61.70	60.18	84.78	14.23	70.80	161	128
ATRC[30]	HRNet18	67.67	62.93	82.29	14.24	72.42	216	96
PAD-Net*[18]	ViT-large	78.01	67.12	79.21	14.37	72.60	773	330
MTI-Net*[17]	ViT-large	78.31	67.40	84.75	14.67	73.00	774	851
ATRC*[30]	ViT-large	77.11	66.84	81.20	14.23	72.10	871	340
InvPT[19]	ViT-large	79.03	67.61	84.81	14.15	73.00	669	423
TaskPrompter[81]	ViT-large	80.89	68.89	84.83	13.72	73.50	497	401
TaskExpert[15]	ViT-large	80.64	69.42	84.87	13.56	73.30	622	420
本章方法	ViT-large	**81.41**	**70.52**	**84.90**	**13.51**	**75.42**	571	407

表 3.8 不同方法在 NYUD-v2 数据集上的定量比较。本章的方法在所有四项任务上表现最佳。

方法	骨干网络	**Semseg** mIoU ↑	**Depth** RMSE ↓	**Normal** mErr ↓	**Boundary** odsF ↑
PAD-Net[18]	HRNet18	36.61	0.6246	20.88	76.38
MTI-Net[17]	HRNet48	45.97	0.5365	20.27	77.86
ATRC[30]	HRNet48	46.33	0.5363	20.18	77.94
InvPT[19]	ViT-large	53.56	0.5183	19.04	78.10
TaskPrompter[81]	ViT-large	55.30	0.5152	18.47	78.20
TaskExpert[15]	ViT-large	55.35	0.5157	18.54	78.40
本章方法	ViT-large	**55.96**	**0.5076**	**18.33**	**78.43**

析和对象边界三个任务的检测性能上分别显著提高了 +0.77 mIoU、+1.10 mIoU 和 +2.12 odsF，同时使用了更少的参数和 FLOPs。

本章还在图 3.6 中直观展示了与其他方法的对比。MLoRE 在语义分割、人体解析和对象边界检测任务上的可视化结果优于之前的 SOTA 方法。

四、 高效的多任务学习模型

本章还将 MLoRE 模块应用于 ViT-small 骨干网络，检查高效模型的性能。具体设置上，解码器的通道数也被从 384 减少到 192。实验结果如表 3.9所示, 本章的方法使用约 TaskExpert 35% 的 GFLOPs，能够实现极具竞争力的结果。特别是在语义分割和目标边界任务上，分别提升了 0.6% mIoU 和 1.01% odsF，而其他任务的指标与 TaskExpert 接近。此外，参数数量比 TaskExpert 少了 11M。

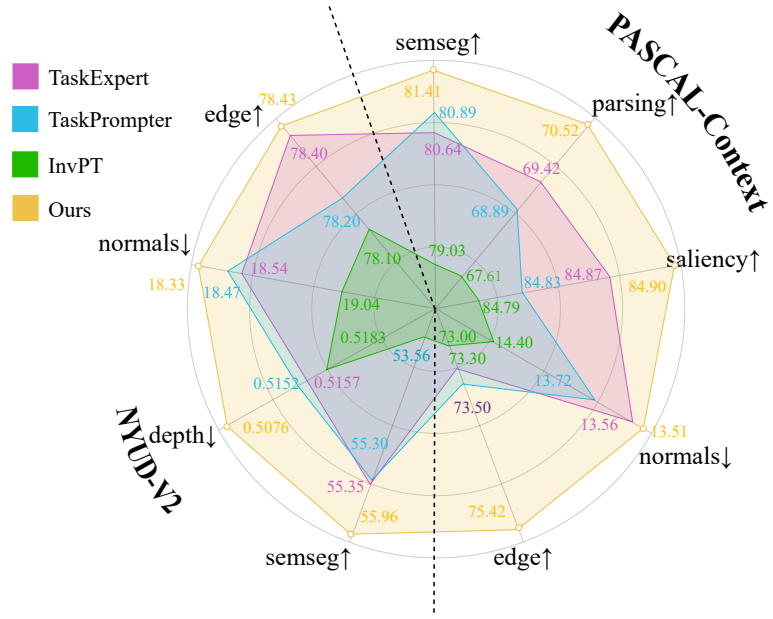

图 3.5 与最先进方法的性能比较。本章基于提出的低秩混合专家模型 MLoRE 在所有任务上都取得了优异的表现。semseg 代表语义分割任务，parsing 代表人体解析任务，saliency 代表显著性物体检测任务，normals 代表法线检测任务，edge 代表边缘检测任务，depth 代表深度估计任务。↑ 表示数值越高越好。↓ 表示数值越低越好。

表 3.9 基于 MoE 的高效模型在 PASCAL-Context 数据集上的定量比较。

方法	**Semseg** mIoU ↑	**Parsing** mIoU ↑	**Sal.** maxF ↑	**Nor.** mErr ↓	**Bound.** odsF ↑	**FLOPs** (G)	# 参数量 (M)
M^3ViT	72.80	62.10	66.30	14.50	71.70	420	42
Mod-Squad	74.10	62.70	66.90	13.70	72.00	420	52
TaskExpert	75.04	62.68	84.68	14.22	68.80	204	55
本章方法	75.64	62.65	84.70	14.43	69.81	72	44

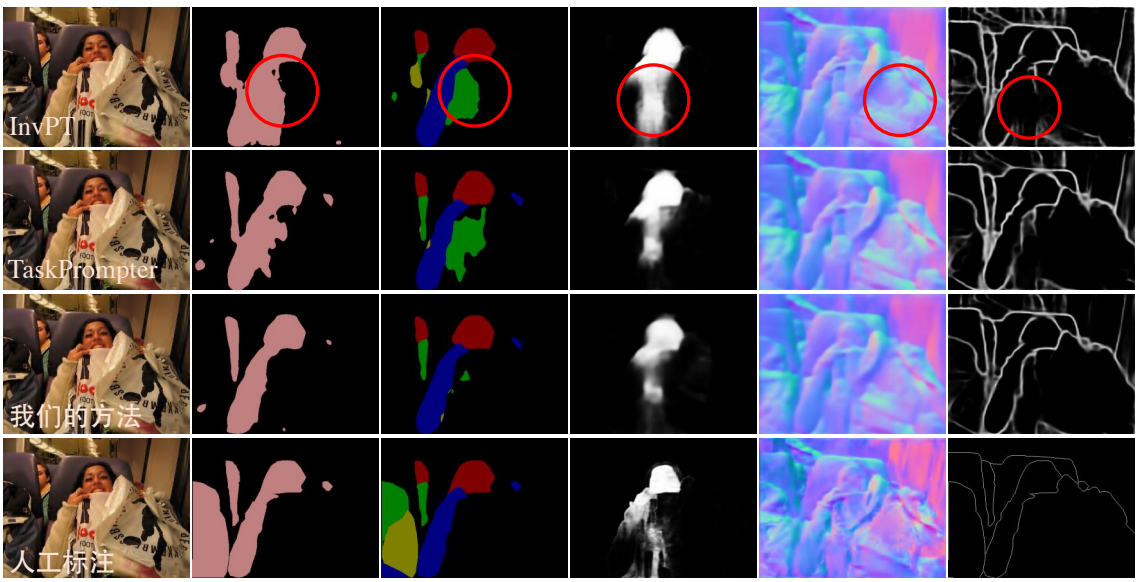

图 3.6 不同方法的定性比较，包括 InvPT[19]、TaskPrompter[81] 和本章所提出的 MLoRE。放大查看效果更佳。可以看到，由于提出的 MLoRE 模块，MLoRE 在五个任务上都取得了比其他方法更好的视觉效果。

第四节 本章小结

本章提出了一种新颖的以解码器为中心的多任务学习方法 MLoRE，通过对标准混合专家模型（MoE）技术的深入分析，从两个维度进行改进以适应密集预测多任务学习需求。首先，针对 MoE 对全局关系建模的忽视问题，本章在 MoE 中引入通用卷积路径，使多任务特征可共享这一路径并建立所有任务之间的关系。其次，本章将标准卷积具有低秩约束的形式应用于不同的专家网络，有效降低专家数量增加带来的计算开销和参数量。实验结果表明，提出的方法在所有指标上明显优于现有的最先进方法，同时还保证了较高的效率，充分验证了本方法的有效性。

第四章　基于扩散模型的多任务密集预测

第一节　背景

如上一章所阐述的，模型设计是多任务密集预测中十分重要的环节，但是由于任务与任务之间在判别式建模方式下存在的固有差异，所以统一的模型设计会不可避免地对模型的性能产生较大负面影响。因此，近来许多通用视觉模型[22, 23, 27, 43, 55, 121] 致力于从建模方式的角度上构建统一架构来处理多类视觉任务。其中，许多方法基于生成式模型取得了突出的效果，如一些通用视觉模型基于扩散模型[24, 25, 28] 将各类密集预测任务重构为统一的标签去噪过程，在不同的任务上都展现出强大的任务处理能力。扩散过程[24, 25, 28] 包含前向加噪过程和反向去噪过程。前向加噪过程逐步向数据样本添加噪声，生成带噪声样本 z_t，可被公式化地表述为：

$$z_t = \sqrt{\gamma(t)} z_0 + \sqrt{1-\gamma(t)}\epsilon, \tag{4.1}$$

其中 ϵ 是高斯噪声，$t \in \{0，1，...，T\}$ 表示时间步。$\gamma(t)$ 是控制信噪比和噪声腐蚀程度的单调递减函数。在前向加噪过程中，原始数据 z_0 被多次迭代破坏从而逐渐趋近于纯高斯噪声 z_T。训练阶段中，通过 θ 参数化的去噪网络 $f_\theta(z,t)$ 通过最小化目标函数（通常采用 l_2 损失函数）学习从 z_t 预测 z_0。在推理阶段，扩散模型执行反向去噪过程。神经网络遵循马尔可夫链（Markovian）方式，从纯高斯噪声 z_T 出发，通过迭代恢复原始数据 z_0。具体而言，$z_T \to z_{T-\delta} \to ... \to z_0$ 的转换过程通过以下步骤实现：对当前噪声状态 z_t 应用去噪网络，利用预测的 $\tilde{z_0}$ 逐步过渡至 $z_{t-\delta}$ 状态。

在基于扩散模型的通用方法中，DDP[27] 通过编解码器解耦设计，将迭代去噪过程限定于解码阶段，实现推理效率提升。相比判别式方法，生成式方法能够显式建模预测目标的条件概率分布[21]，在图像细节部分的预测上更具优势。具体而言，在感知任务[27, 43] 中，扩散模型通常以特征 $\mathbf{x}$ 为条件进行去噪。例如在语义分割任务中，扩散模型将带噪声的分割标签 z_t 与条件特征 $\mathbf{x}$ 共同作为输入，

执行去噪过程。该条件扩散过程被公式化地表述如下：

$$q_\theta(z_{0:T}|\mathbf{x}) = q(z_T)\prod_{t=0}^{T} q_\theta(z_{t-1}|z_t,\ \mathbf{x}), \tag{4.2}$$

其中 $q_\theta(\cdot)$ 通过基于去噪网络 $f_\theta(z,\ t,\ \mathbf{x})$ 的转移规则实现，该网络以 $\mathbf{x}$ 作为条件输入。

尽管通用模型在不同任务中显示出优于定制模型的优势，但在实际应用中，如自动驾驶和虚拟现实领域，感知模型通常需要在一系列密集预测任务中进行推理。在这些情况下，以往为每个任务单独训练一个模型的通用方法需要进行多次前向推理才能生成所有任务的预测结果，这使得推理阶段的效率较低。此外，尽管扩散模型已被证明能够捕捉每个单一任务的潜在分布，但其捕捉跨任务关系的潜力仍有待发掘。跨任务关系是提升多任务框架中不同密集预测任务整体性能的关键[19, 30]，这些因素促使本章探索基于扩散模型的通用模型是否能够扩展到多任务密集预测领域，尤其是考虑到其在处理各种密集预测任务中的潜力[27, 43, 44]。

直接将扩散模型应用于多任务密集预测面临几个明显的挑战。首先，为多个任务分别进行去噪会阻碍扩散模型挖掘任务之间的关系。此外，不同任务的目标标签具有异质性（例如，语义分割的离散类别标签和深度估计的连续标签）。这需要为不同任务的标签设计繁琐的任务特定编码（例如，为离散标签设计的模拟比特[103]）。最后，扩散模型通过迭代去噪过程生成最终预测，这需要多次前向传递才能为每个任务输出最终结果。在处理多个任务时，为每个任务执行多次前向推理会导致效率降低。

针对这些挑战，本章提出了一种新颖的多任务扩散网络，命名为 TaskDiffusion。本章的 TaskDiffusion 将不同任务的去噪扩散过程耦合到解码器中的一个联合去噪扩散过程中。具体来说，本章的联合去噪扩散过程包括跨任务标签编码和跨任务扩散解码器。对于跨任务标签编码，本章使用嵌入层来编码不同任务标签，并将这些特征的拼接映射为跨任务特征图。这种编码策略可以在不使用复杂任务专用编码方法的情况下，转换来自不同任务的异构标签。对于跨任务扩散解码器，本章采用基于从不同层级提取的任务专用特征进行条件约束的跨任务扩散解码器。与先前工作[27] 中为不同任务分别使用专用去噪解码器的做法不同，本章的 TaskDiffusion 通过显式建模任务间关系及层级间关系来执行扩散过程，这在多任务学习场景[17] 中至关重要。本章在图 4.1中展示了本章方法

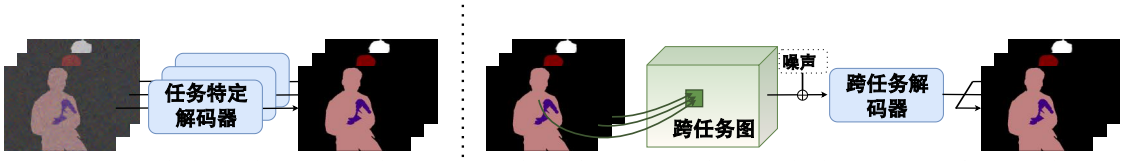

图 4.1 左图为任务专用扩散过程，右图为本章提出的联合扩散过程对比示意图。本方法将不同任务的标签编码为统一的跨任务特征图，并采用单一跨任务扩散解码器完成去噪处理。

与任务专用扩散过程的简要对比。

本章是率先将扩散模型应用于全标注多任务密集预测的研究之一，且本方法相较此前方法能够取得显著性能提升。为验证方法的有效性，本章在 PASCAL-Context 和 NYUD-v2 数据集上开展了全面实验。实验表明，本方法在所有任务上的性能均超越了此前最先进方法。通过揭示扩散模型在该领域的有效性，本章相信基于扩散的方法仍有巨大潜力值得进一步探索。本章希望本方法能为相关研究领域带来新的启示。综上所述，本章的贡献包含以下三个方面：

- 本章探索了如何利用扩散模型作为多任务密集预测的有效求解方法，并提出一种新颖的联合去噪扩散过程来捕捉任务间的关系。
- 本章提出跨任务标签编码策略以摒弃复杂的任务专用编码方法，并设计跨任务扩散解码器来显式建模任务间关系及层级间关系。
- 本章在 PASCAL-Context 和 NYUD-v2 基准数据集上进行了大量实验。实验结果表明，TaskDiffusion 在所有任务上的性能均优于此前最先进的方法。

第二节 方法

一、 结构

本章提出的 TaskDiffusion 整体框架如图 4.2所示。本章的整个框架由像素级编码器、跨任务标签编码器和跨任务扩散解码器三个核心组件构成。

像素级编码器：像素级编码器以图像 $\mathbf{I}$ 为输入，为每个任务提取多层级特征集合 $\{\mathcal{F}_i^s \in \mathbb{R}^{C\times H\times W} | i \in \{1，2，...，N\}, s \in \{1，2，...，S\}\}$，下文将其统一记为 $\{\mathcal{F}_i^s\}$ 。H, W 和 C 分别表示单层级特征图的高度、宽度和通道维度。N 代表特征层级总数，S 表示任务总数。具体而言，本章延续先前多任务方法[19, 81], 的设计范式，首先使用共享的 Vision Transformer（ViT）主干网络[73] 为所有任务提取任务通用特征。本章选取主干网络不同层级的特征，记为 $\{\mathbf{X}^l | l \in \{l_1，l_2，...，l_N\}\}$。其中 l 表示层索引，$\mathbf{X}^l$ 代表从主干网络第 l 个选定

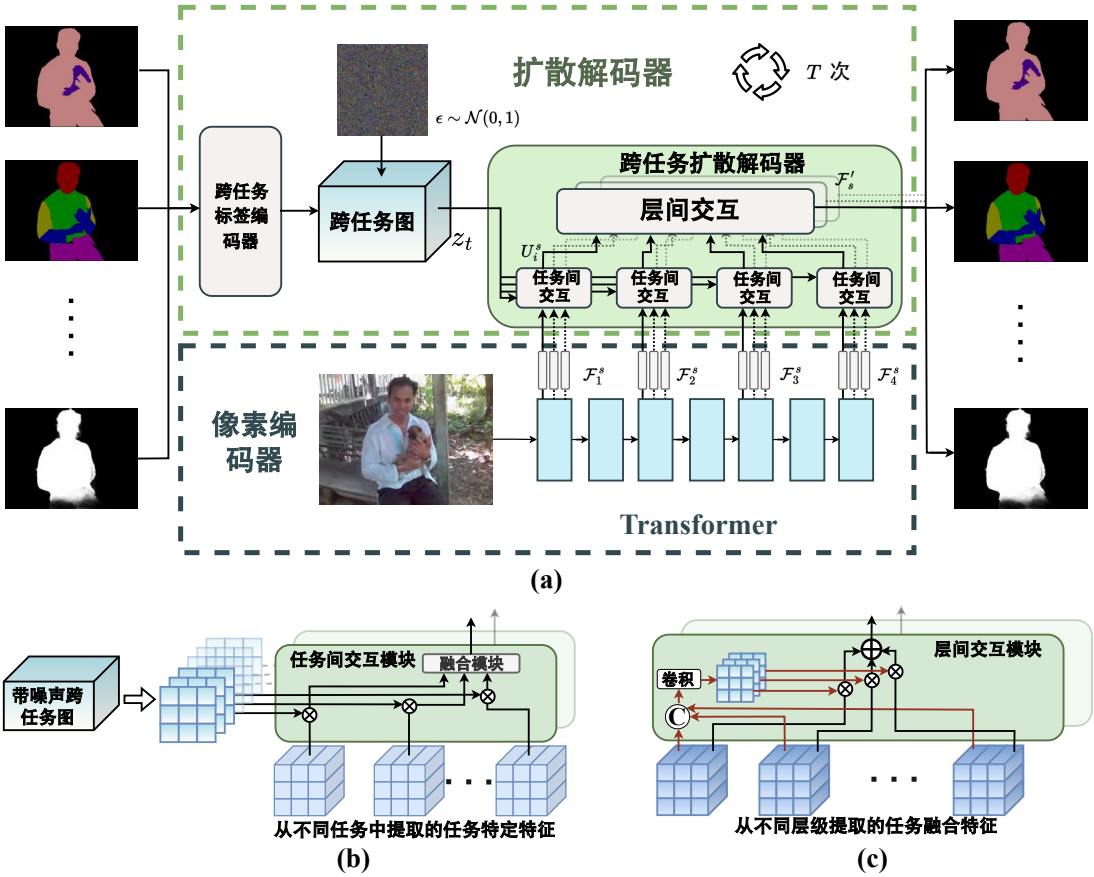

图 4.2 a) 所提方法的整体框架。跨任务扩散解码器以任务专用的多层级特征作为条件，对带噪声的跨任务特征图执行迭代去噪。该解码器由任务交互模块和层级交互模块构成，可显式建模任务间关系与层级间关系，并利用这些关系融合来自不同任务、不同层级的任务专用特征。聚合后的特征用于预测不同任务，各任务的预测逻辑输出被送入跨任务标签编码器，生成预测的跨任务特征图并执行迭代推理。b) 任务交互模块结构。c) 层级交互模块结构。

层级提取的特征。将这 N 个层级的特征输入至 S 个任务专属分支中，生成任务特定的多层级特征 $\{\mathcal{F}_i^s\}$。每个任务专属分支包含两个堆叠的卷积模块，每个模块由 3×3 卷积层、批归一化层、GeLU 激活层以及 1×1 卷积层依次连接构成。

本章通过设计任务专属辅助头（task-specific auxiliary heads）来生成中间预测结果，以此学习不同任务间具有判别性的多层级任务特定特征。这些中间预测通过对应任务的标注数据 $\{\mathbf{K}_s\}$ 进行监督训练，任务专属分支将通过来自各自任务的梯度进行参数更新。这种做法在多任务密集预测中属于常规操作，先前研究[17, 19, 30]. 都有类似的训练方式。每个任务专属分支的输出将作为跨任务扩散解码器的条件输入。

跨任务标签编码器：在多任务密集预测中，模型需要学习具有异构标签的多个任务，例如离散的分类标签和连续的深度标签。然而，如先前研究[43, 103]所示，扩散过程难以有效处理离散标签。因此，这类标签需要单独进行预处理[27, 103]。

然而，若将不同任务的标签编码至不同的特征空间，则会阻碍模型捕捉任务间的关联性，而这正是多任务学习的关键要素[19, 30]。此外，为不同任务设计专用编码方法既繁琐又难以推广到新任务。基于此，本章提出跨任务标签编码机制，将异构标签统一映射至联合连续特征空间。该编码机制还能帮助扩散解码器在单次前向过程中捕获任务间关联性。

本章所提出方法首先通过统一的标签编码器将不同任务标签 $\{\mathbf{K}_s|s \in \{1，2，...，S\}\}$ 映射至特征空间。对于离散标签（如语义分割标签），本章所提出方法将其转换为独热编码形式对于连续标签，则直接将原始标签输入编码器。该标签编码器包含一个 1×1 卷积层，将任务标签映射为任务专属的编码标签 $\{\mathbf{K}'_s \in \mathbb{R}^{C\times H\times W}|s \in \{1，2，...，S\}\}$。

为进一步捕捉任务间关联性，本章所提出方法将不同任务的编码标签进行通道维度拼接，并通过另一个 1×1 卷积层将其映射为跨任务特征图 z。此联合映射过程综合考虑所有编码特征图，在高维特征空间中对任务关系进行建模。遵循先前方法[27, 43]，本章所提出方法将跨任务特征图 z 归一化至 $[-1，+1]$ 区间，并通过缩放因子 *scale* 控制信号噪声比。该设计能有效增强去噪任务与扩散解码器的训练难度。

在将不同任务标签编码为跨任务特征图 z 后，高斯噪声将被注入以生成被破坏掩码 z_t。如式 4.1所示，$\gamma(t) \in [0，1]$ 控制着噪声的强度，该噪声的强度随时间步 t 递增而递减。本章延续先前工作[27] 的设置，采用余弦噪声调度策略[122]对 $\gamma(t)$ 进行时序控制。

跨任务扩散解码器：解码器以带噪的跨任务特征图 $z_t \in \mathbb{R}^{C\times H\times W}$ 作为输入，并以任务特定的多层级特征 $\{\mathcal{F}_i^s\}$ 作为条件。本章同时引入任务共享特征 $\mathbf{X}^l$ 以辅助建模跨任务关联性。带噪标签图 z_t 与 $\mathbf{X}^l$ 沿通道维度拼接后输入跨任务扩散解码器，通过 $\{\mathcal{F}_i^s\}$ 依次执行任务间交互和层级间交互。

具体而言，在任务交互阶段，TaskDiffusion 在每个特征层级上进行任务交互。本章所提出方法使用 2 个卷积模块（每个块包含卷积-批归一化-GeLU 激活-卷积的级联结构）和一个卷积层，将拼接后的跨任务带噪图 z_t 与 $\mathbf{X}^l$ 映射至各特征层级，并将通道数从 C 变换为 S^2。通过对每个层级的卷积模块输出特

Algorithm 1 TaskDiffusion 训练阶段

```
def train(images, masks_gts):
  """
  images: [B, 3, H, W]
  masks_gts: {task:[B, *, H, W]}
  tasks: 需要预测的所有任务列表
  """
  # 编码图像生成任务特定的多层级特征
  feats =pixel_encoder(images)

  # 为每个任务编码对应的真实掩码
  for task in tasks:
    m_enc[task] =ts_label_encoder[task](
                                masks_gts[task])

  # 将所有任务编码结果融合成跨任务特征图
  m_enc_cross =label_encoder(cat(m_enc)) *scale

  # 根据时间步对跨任务图添加噪声（扩散过程）
  t =randint(0, T) # timestep
  eps =normal(mean=0, std=1) # 标准高斯噪声
  m_crpt =sqrt(alpha_cumprod(t)) *m_enc_cross +
             sqrt(1 -alpha_cumprod(t)) *eps

  # 通过扩散解码器进行预测
  m_preds =diff_decoder(m_crpt, feats, t)

  # 计算各任务的预测损失
  for task in tasks:
    loss[tasks] =prediction_loss[tasks](m_preds[
                                task], masks_gts)

  return loss
```

Algorithm 2 TaskDiffusion 推理阶段

```
def infer(images, steps):
  """
  images: [B, 3, H, W]
  steps: 逆向扩散过程的采样步数
  """

  # 提取各任务图像的多层级特征
  feats =pixel_encoder(images)

  # 初始化随机噪声作为初始跨任务图
  m_t =normal(mean=0, std=1)

  for step in range(steps):
    # 非对称时间步计算
    t_now =1 -step / steps
    t_next =max(1 -(1 +step +t_diff) / steps, 0)
                              # 非对称时间采样

    # 通过扩散解码器预测去噪结果
    m_preds =diff_decoder(m_t, feats, t_now)

    # 对各任务预测结果进行编码
    for task in tasks:
        m_enc[task] =ts_label_encoder[task](
                                  m_preds[task])
    # 编码生成融合后的跨任务特征图
    m_enc_cross =label_encoder(cat(m_enc)) *
                                scale
    # 估计t_next时刻的噪声图
    m_t =ddim(m_t, m_enc_cross, t_now, t_next)

  return m_preds
```

征施加 Sigmoid 函数，生成任务关系图 $\{\mathbf{A}_i \in \mathbb{R}^{S\times S\times H\times W} | i \in \{1，2，...，N\}\}$。任务融合特征通过将 $\{\mathbf{A}_i\}$ 与任务特定的多层级特征 $\{\mathcal{F}_i^s\}$ 进行矩阵乘积累加生成，其数学表达为 $\mathbf{U}_i^s = \sum_{p=1}^{S} \mathbf{A}_i^{s,p} \cdot \mathcal{F}_i^p$。随后，这些任务融合特征被输入至层级交互阶段。

在层级交互阶段针对任务 s 来自不同层级的任务融合特征 $\{\mathbf{U}_i^s \in \mathbb{R}^{C\times H\times W} | i \in \{1，2，...，N\}\}$，本章沿通道维度进行拼接后输入另一个卷积模块，该卷积模块将通道数从 $N \times C$ 映射至 N。通过 Sigmoid 函数对输出特征进行处理，生成任务特定的层级融合图 $\mathbf{M}^s \in \mathbb{R}^{N\times H\times W}$。

本章所提出方法使用 $\mathbf{M}^s$ 通过公式 $\mathcal{F}'_s = \sum_{i=1}^{N} \mathbf{M}_i^s \cdot \mathbf{U}_i^s$ 生成聚合后的任务特定特征，其中 $\mathcal{F}'_s$ 表示聚合任务特定特征，$\mathbf{M}_i^s$ 表示该张量第一个维度的第 i 个元素。聚合后的任务特定特征集合 $\{\mathcal{F}'_s | s \in 1，2，...，S\}$ 将被输入到各任务对应的预测分支中生成最终预测结果。每个任务特定预测分支包含三个卷积模块组成的结构。生成的预测结果会通过前文所述的编码方式处理，最终输出预测的跨任务特征图 z_0。此外，本章将在后续的小节中详细分析跨任务扩散解码器的设计优势。

二、 训练与推理

在训练阶段，本章所提出方法首先生成跨任务特征图 z 并添加噪声，获得带噪特征图 z_t。该带噪特征图被输入到扩散解码器中，随后训练模型执行去噪过程，完整训练流程如算法 1所示。在推理阶段，扩散模型以高斯噪声为起点，通过迭代去噪使特征图逐步逼近不同任务的真实分布。该过程总结于算法 2中。

损失函数：在图像生成任务的扩散模型[25, 28] 中，通常采用 l_2 损失函数。但面向感知任务的扩散模型改进方法[23, 43] 表明，判别性损失函数较标准 l_2 损失具有更优性能。与上述方法不同，本章方法涉及多任务联合训练，各任务采用差异化损失函数。因此，本章不再使用 l_2 损失监督预测结果 z_0。而是通过加权任务专属损失（如语义分割任务采用交叉熵损失，深度估计任务采用 l_2 损失）对不同任务的预测逻辑值进行监督学习。具体而言，本方法采用的损失函数可表示为：

$$\mathcal{L}_{all} = \sum_{s=1}^{S} w_s \mathcal{L}_s(k_s,\ K_s), \tag{4.3}$$

其中 k_s 表示任务 s 的预测逻辑值，w_s 为任务 s 的损失权重系数，$\mathcal{L}_s$ 对应任务 s 的专属损失函数。本章沿用了[81] 中针对不同任务的损失函数设计策略，具体实现细节将在 第三节实验设置部分详细阐述。

联合去噪：在多任务密集预测领域，基于解码器的经典方法[19, 81] 通常采用共享编码器配合多个任务专属解码器的架构来生成各任务预测。此外，构建任务关联时多层次特征交互也至关重要[17].[17]。由此容易联想到基于多层级任务通用特征构建多个任务专属扩散解码器的设计思路。但该设计会导致模型性能与效率的双重下降，原因有二：其一，解码器需要同时学习任务专属特征、任务通用特征以及任务关联关系[81]，而任务专属扩散解码器难以有效建模任务通用特征与任务关联关系。作为对比，本章方法通过在跨任务扩散解码器中实施如图 4.4所示的由粗到细的任务关联建模，使得迭代去噪过程能够生成更精确的预测结果。其二，扩散解码器需通过多次前向传播生成最终预测。通过多任务联合去噪，本方法不仅实现了解码器内的任务关联建模，更在推理阶段显著降低了计算开销（时间复杂度分析详见附录）。

联合去噪时间复杂度对比分析：本章通过对比任务专用扩散解码器与跨任务扩散解码器的时间复杂度，验证所提架构的计算效率。以基于多层次骨干特征 $\{\mathbf{X}^l\}$ 构建的多层次任务专用解码器为例，假设每个任务在不同层级包含两个

堆叠的卷积模块。单次前向传播中这些卷积模块的总时间复杂度为 $O(2NS)$。其中 N 表示层级数量，S 表示任务总数。各层级特征首先与隐变量 z_t 沿通道维度拼接，并通过这两个卷积模块进行处理。各层级输出特征经拼接后，通过 1×1 卷积层将通道数从 $4C$ 映射至 C。该过程的时间复杂度在简化分析时可忽略不计。映射后的特征再经过 3 个堆叠卷积模块生成最终 logits，对应时间复杂度为 $O(3S)$。推理阶段，$\{\mathbf{X}^l\}$ 中各特征均与相同 z_t 沿通道维度拼接，不同层级处理结果输入解码器生成去噪后的 $z_{t-\delta}$。该任务专用扩散解码器设计在生成 z_0 时的总体时间复杂度为 $O(T(2NS+3S))$。

针对本章的联合去噪方法，首先生成任务特定的多层次特征所需时间复杂度为 $O(2NS)$。在每次迭代过程中，跨任务图 z_t 需经过 2 个卷积模块和各层级的卷积操作来生成任务关系图，对应时间复杂度 $O(2N)$。如 **??**所述，生成的任务关系图将用于构建任务特定的多层次特征及聚合任务特征。任务交互与层级交互的时间复杂度在简化分析时可忽略不计。聚合后的任务特征还需通过 3 个堆叠卷积模块生成最终 logits，该过程时间复杂度为 $O(3S)$。综上，本章的联合去噪方法总体复杂度为 $O(2NS+2NT+3ST)$，通过多层级卷积模块的并行前向传播显著降低了计算开销。具体而言，本章的解码器在 3 次前向传播中 FLOPs 为 568G ，相比任务专用扩散解码器的 779G 具有明显优势。

采样策略：在本章方法中，本章采用 DDIM[28] 作为特征图更新规则。当每个时间步预测得到 z_0 后，通过重参数化技巧生成下一步的带噪跨任务特征图。参照[27, 43, 103] 的设计，本章在推理阶段采用非对称时间间隔策略。该时间间隔由算法 2中的 t_diff 参数控制，经验性设定其取值为 1。

第三节 实验

一、 实验设置

数据集与评估指标：本章在两个公开多任务数据集上进行实验评估，包括 PASCAL-Context[118] 和 NYUD-v2[119]。PASCAL-Context 数据集包含 4,998 张训练图像和 5,105 张测试图像，提供五项密集预测任务的标注：语义分割、人体解析、显著性检测、表面法线预测和边界检测。表面法线预测和显著性检测的标签来自先前工作[120]。NYUD-v2 数据集包含 795 张训练图像和 654 张测试图像，提供四项密集预测任务的标注：语义分割、单目深度估计、表面法线预测和边界检测。两个数据集的输入分辨率分别为 512×512 和 448×576。在后续实验结果

中，本章用 semseg 代表语义分割任务，parsing 代表人体解析任务，saliency 或者 sal. 代表显著性物体检测任务，normals 代表法线检测任务，edge 或者 Bound. 代表边缘检测任务，depth 代表深度估计任务。

遵循先前工作[15, 19, 81], ，本章使用平均交并比（mIoU）评估语义分割任务和人体解析任务，使用均方根误差（RMSE）评估单目深度估计任务，使用平均角度误差（mErr）评估表面法线预测任务，并通过最优数据集尺度 F 值（odsF）评估边界检测任务。为综合评估所有任务的整体性能，本章采用[120] 中的多任务学习增益指标 (Δ_m)，该指标通过加权各任务相对于单任务基线的性能增益来衡量多任务协同效果。

训练与推理细节：遵循先前工作[15, 19, 81], ，本章使用 ViT-large[73] 作为主干网络，所有消融实验的主干网络均采用 ViT-base 。训练批大小设置为 4 ，所有实验均训练 40000 次迭代。初始学习率在 PASCAL-Context 设为 2e-5，在 NYUD-v2 设为 1e-5，两个数据集均采用 1e-6 的权重衰减。本章沿袭先前方法[81] 采用多项式学习率调度器。对于具有连续标签的任务（如深度估计和表面法线预测），本章使用 l_1 损失函数；对于具有离散标签的任务（如语义分割、人体解析、显著性目标检测和边界检测），则采用交叉熵损失函数。参照先前工作[81] 设置损失权重来实现不同训练损失间的平衡。在 PASCAL-Context 数据集上，本章使用独立卷积层编码显著性目标检测标签，而其他四个任务则通过跨任务标签编码器进行特征编码。最终跨任务特征图 z 由这两个编码特征的拼接构成。所有实验均在 2 块 NVIDIA V100 GPU 上完成 40000 次迭代训练。

二、 与最先进方法的对比

本章在 PASCAL-Context 和 NYUDv2 数据集上将提出方法与现有最先进方法进行定量对比。如表 4.1 和 表 4.2所示，本章的方法在两项数据集的所有任务上均显著优于多数先前方法。由于提出的 TaskDiffusion 是首个利用扩散模型显式建模任务关系的新框架，其可与现有多任务架构协同实现性能提升。通过与 MLoRE[126] 方法结合，本章的方案在 PASCAL-Context 数据集上以 Δ_m 指标超越 MLoRE[126] 和 TSP-Transformer[125] 分别达 +0.97% 和 +0.58% 。 本章的方法在推理过程中虽需迭代执行去噪步骤，但与 TaskExpert[15] 相比仍保持竞争力。(610GFLOPs vs 622 GFLOPs) 此外，本章在全标注条件下复现了另一基于扩散模型的多任务密集预测方法 DiffusionMTL[110]。实验表明，在参数量相当的条件下，本章方法在性能与计算成本两方面均明显优于 DiffusionMTL。这些实验结

表 4.1 PASCAL-Context dataset 数据集不同方法的定量对比。† 表示由本章复现的基于 ViT-large 主干网络的方法。 * 表示基于[15] 中 ViT-large 主干网络复现的方法。** 表示由本章在全标注条件下复现的方法。本章的方法在所有五项任务中均表现最佳。↑表示数值越高越好。↓表示数值越低越好。

方法	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	Δ_m % ↑	FLOPs (G)	# 参数量 (M)
单任务学习	81.62	72.21	84.34	13.59	76.79	-	-	-
MTI-Net*[17]	78.31	67.40	84.75	14.67	73.00	-4.62	774	851
ATRC*[30]	77.11	66.84	81.20	14.23	72.10	-5.50	871	340
MQTransformer†[123]	77.72	65.14	84.43	14.63	54.77	-10.16	360	314
DeMT†[124]	78.96	67.39	84.26	14.53	55.29	-8.99	372	308
InvPT[19]	79.03	67.61	84.81	14.15	73.00	-3.61	669	423
TaskPrompter[81]	80.89	68.89	84.83	13.72	73.50	-2.03	497	401
TaskExpert[15]	80.64	69.42	84.87	13.56	73.30	-1.74	622	420
DiffusionMTL**[110]	80.46	69.13	84.85	14.02	70.96	-3.16	732	381
TSP-Transformer[125]	81.48	70.64	84.86	13.69	74.8	-1.01	1991	422
MLoRE[126]	81.41	70.52	84.90	13.51	75.42	-0.62	571	407
本章方法	81.21	69.62	84.94	13.55	74.89	-1.11	610	416
本章方法 /w MLoRE	**81.58**	**71.3**	**85.05**	**13.43**	**76.07**	**-0.04**	738	472

表 4.2 NYUD-v2 dataset 数据集不同方法的定量对比。所有的方法均基于 ViT-large 骨干网络

方法	Semseg mIoU ↑	Depth RMSE ↓	Normal mErr ↓	Boundary odsF ↑	Δ_m % ↑
单任务学习	56.77	0.5141	18.56	78.93	-
InvPT[19]	53.56	0.5183	19.04	78.10	-2.52
TaskPrompter[81]	55.30	0.5152	18.47	78.20	-0.81
TaskExpert[15]	55.35	0.5157	18.54	78.40	-0.84
TSP-Transformer[125]	55.39	0.4961	18.44	77.5	-0.02
MLoRE[127]	55.96	0.5076	18.33	78.43	0.11
本章方法	55.65	0.5020	18.43	78.64	0.18
本章方法 /w MLoRE	56.66	0.5033	18.13	78.89	1.04

果验证了所提出联合去噪扩散过程的有效性。可视化预测结果详见附录。相较于先前最先进方法 TaskPrompter，本章方法在语义分割、人体解析及边界检测等任务上能生成更优的预测结果。同时，如表中所示，本章的方法与前一章所提出的 MLoRE 结合之后能实现更加出色的效果，具体来说，其性能在大部分任务上相比单独的 MLoRE 或者 TaskDiffusion 都要更高。这一结果同样表明本章所提出的两种方法在各自专注的方面取得了有效的改进。

此外，为直观对比不同方法的性能差异，本章在图 4.3中展示了可视化结果。可以看出本章的方法在所有任务上均取得了更精确的预测结果。

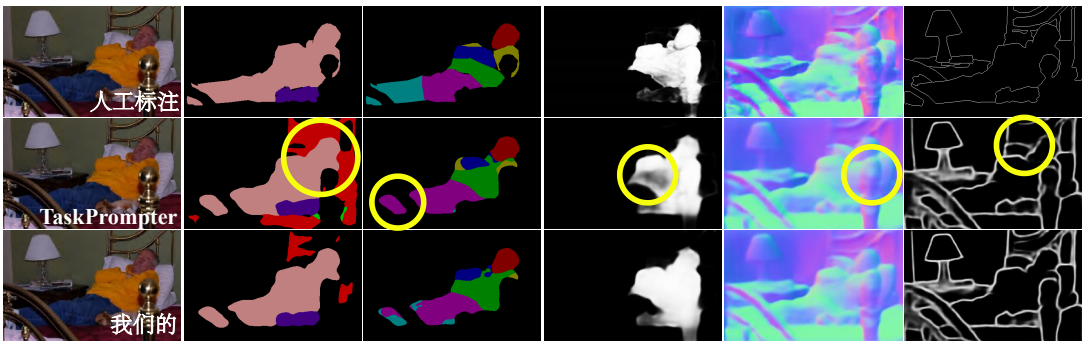

图 4.3 定性对比实验，与先前最佳方法[81] 的对比结果。建议放大来查看细节。可以看出本章的预测取得了更优的结果。

表 4.3 不同组件的消融实验。

实验设置	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	Δ_m % ↑
单任务学习	79.63	69.76	85.37	13.41	76.15	-
基线	77.34	66.17	85.18	13.78	72.40	-2.80
w/ 任务特定解码器	77.41	67.15	85.29	13.83	74.40	-2.02
w/ 跨任务扩散解码器	78.51	67.32	85.26	13.47	74.60	-1.11
+ 跨任务标签编码	78.83	67.40	85.31	13.38	74.68	-0.84

三、 消融实验

消融实验的基准模型采用 ViT-base 作为主干网络，直接从 ViT 的第 3, 6, 9 和 12 层提取多层级特征。这些不同层级的特征经拼接后输入任务特定分支（包含 1 个 1×1 卷积层和 2 个卷积模块），生成融合特征 F_s^{fused}。该融合特征用于生成最终预测结果。所有消融实验均在 PASCAL-Context 数据集进行，默认执行 3 次去噪迭代完成推理预测。缩放因子 *scale* 默认设置为 0.01（特殊说明情况除外）。更多消融分析结果可参见附录部分。

不同组件的有效性分析：本章通过消融实验验证联合去噪扩散过程中各组件的作用效果，详见表 4.3。首先测试融合特征 F_s^{fused} 驱动的任务特定扩散解码器效果，实验表明该组件的引入使多任务学习增益（MTL gain）显著提升。当将普通任务扩散解码器替换为提出的跨任务扩散解码器后，所有任务性能均得到进一步改善，这归因于该模块显式建模了任务间关联与层级关系，同时降低了计算复杂度（附录将展开讨论）。最后，当引入跨任务标签编码机制时，所有任务指标及 MTL 增益均有提升，证实了在标签编码阶段捕获任务关联的重要性。

表 4.4 跨任务扩散解码器中任务交互与层级交互影响的消融实验

实验设置	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	Δ_m % ↑
交叉注意力机制	78.39	67.16	85.42	13.61	74.10	-1.49
特征拼接	78.31	67.58	85.25	13.51	74.80	-1.10
任务交互（本章的方法）	78.83	67.40	85.31	13.38	74.68	-0.84
w/o 层级交互	78.03	67.62	85.27	13.42	74.55	-1.09
w 层级交互	78.83	67.40	85.31	13.38	74.68	-0.84

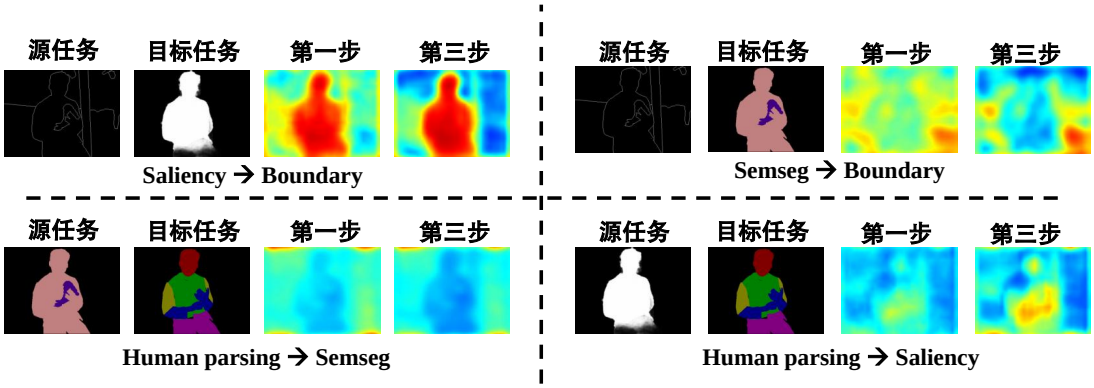

图 4.4 任务关系掩码 A_i 的可视化分析。在每组任务对中，本章展示了第一推理步骤和第三推理步骤中部分任务对的任务关系掩码。可以观察到，第三步骤的任务关系掩码更集中关注特定区域。

任务交互与层级交互的有效性分析：本章通过消融实验验证跨任务扩散解码器中任务交互与层级交互模块的有效性。针对任务交互机制，本章测试了三种不同实现方式。第一种在不同层级应用交叉注意力机制，以跨任务噪声图 z_t 作为查询向量、条件特征作为键值对；第二种直接将条件特征与跨任务特征拼接（跨任务特征通过在各层级使用 2 个卷积模块处理 z_t 生成）；第三种采用 第二节所述的任务交互机机制。

如表 4.4所示。显式建模任务关系的交互机制性能优于交叉注意力与特征拼接方案。关于层级交互，实验表明添加该模块后语义分割 mIoU 从 78.03% 提升至 78.83%，多任务增益 Δ_m 从 -1.09% 改善至 -0.84%。这些实验都验证了层级交互的有效性。

缩放因子：本章在第二节提到的缩放因子 *scale* 上进行了消融实验。实验结果如 表 4.5所示。随着缩放因子的减小，Δ_m 的整体性能相应提升，并在 *scale* 为 0.01 时达到峰值。本章分析认为，当缩放因子增大时，模型更难通过高噪声样本进行去噪训练，这会损害扩散模型的去噪能力。当继续将缩放因子降低至 0.001

表 4.5 缩放因子 *scale* 的消融实验。

scale	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	Δ_m % ↑
0.04	76.47	66.03	85.06	13.45	74.80	-2.01
0.02	78.78	67.25	85.44	13.48	74.83	-0.98
0.01	78.83	67.40	85.31	13.38	74.68	-0.84
0.001	78.72	67.58	85.25	13.51	74.80	-1.00

表 4.6 推理过程的消融实验。

步数	Δ_m % ↑	FLOPs (G)
1	-1.47	508
3	-0.84	667
5	-0.83	827

时，性能未呈现进一步改善。

推理步数：本章对推理步数进行消融实验，结果如 表 4.6所示。随着推理步数的增加，模型性能显著提升。具体而言，所有任务在 3 步推理下的性能均高于 1 步推理。当继续增加推理步数时，性能未见明显改善。此外，更多推理步数会导致计算效率下降。本方法选择 3 步推理，从而在性能与效率之间取得平衡。为直观展示迭代推理的有效性，本章在图 4.4中可视化不同层次任务对的任务关联图 A_i 。可观察到相较于第一步的关联图，第三步的关联图聚焦于更精细的区域。这些结果表明本章的跨任务扩散解码器通过迭代推理实现了由粗到精的优化过程。

避免负面知识转移：多任务学习中的一个关键挑战是任务干扰问题，即所谓的负面知识转移现象。不同任务间的负面知识转移源于接收到的梯度会朝相互冲突的方向更新，导致效果抵消[128]。为解决此问题，本章设计的跨任务扩散解码器通过以下两个机制发挥作用：首先，在解码器中通过显式学习任务关系图建立像素级的任务关联建模。针对每个任务，跨任务扩散解码器能够动态激活对当前任务有益的特征，同时抑制产生冲突的特征。此外，如图 4.4所示，本章的解码器可在不同扩散步中建模差异化的任务关联模式。该方法通过迭代推理实现由粗到精的任务处理流程，有利于生成细粒度的任务关联表征。其次，本章引入了包含多级任务特定分支的模块化架构，该设计可有效缓解负面转移问题[128]。为确保多级任务特定分支生成任务专属特征，本章额外增设辅助预测头对中间特征进行监督。这种辅助损失机制使各任务分支能够获取针对性的梯度更新，该策略在多任务密集预测领域已被广泛验证[17, 19, 110]。

跨任务图的可视化分析：本章可视化标签编码器获得的跨任务图，以展示不同样本在图 4.5中提取的任务信息类型。可以看出，标签编码器生成的跨任务地图能够动态捕捉不同任务的判别性特征。例如在图 4.5左侧，具有不同人体解析标签的像素被编码为不同特征。同时可以观察到，具有不同法线值的像素（如肩部像素和躯干像素）也被编码为不同特征。这验证了本章的跨任务地图编

码方法能够有效整合来自不同任务的判别信息。而在图 4.5右侧，不同人体解析标签像素间的编码特征差异相对较小。相反，具有不同语义标签（人类与犬类）的像素编码特征表现出更显著差异。这表明本章的方法能够根据样本特性动态调整跨任务地图的编码方式。

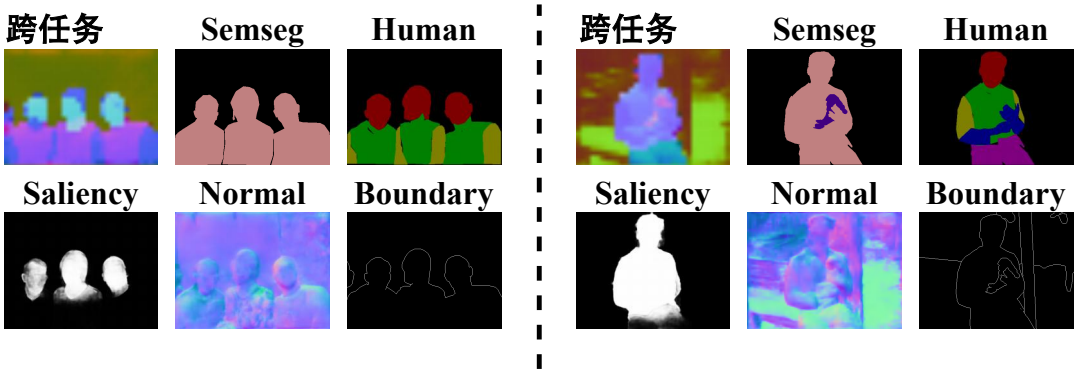

图 4.5 本章的跨任务图及其他任务标签的可视化。通过主成分分析 (PCA) 的跨任务图可视化。

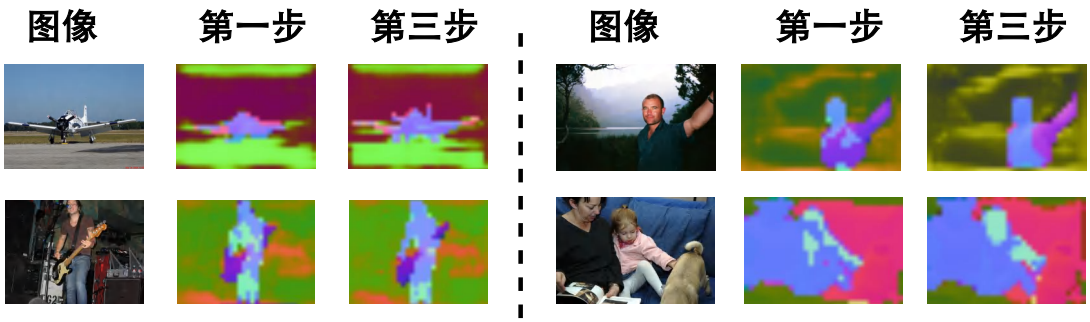

图 4.6 不同步骤跨任务图的可视化。通过主成分分析 (PCA) 的跨任务图可视化。

为更好理解扩散过程，本章还对不同步的跨任务图进行了可视化。从图 4.6中可以看出，第一步的跨任务图相比第三步的跨任务图精度较低且噪声更多。这印证了本章的扩散过程通过交互式推理实现了由粗到精的特征优化。

可视化任务关系掩码的分析： 在图 4.4中，本章针对特定 s, p 和 i 对 第二节提及的 $A_i^{s,p}$ 进行了可视化。本章在每幅子图底部标注了任务对。例如图 4.4左上方子图中可视化的 $A_i^{s,p}$，p 表示显著性检测任务索引，s 表示边界检测任务索引。为更好理解生成的任务关系掩码 A_i，本章将在下文对图 4.4中所有可视化注意力图进行系统分析。

首先，神经网络需要从原始图像中提取多样化特征以生成高质量预测结果。例如，部分特征聚焦于边界信息，而另一些则关注图像主体区域。通过多角度特征的综合分析，神经网络能够在不同任务中实现精确预测。虽然跨任务注意力图与预测结果未必存在直接关联，但其聚焦于能够提升预测精度的关键特征。为利用扩散模型建模任务间关系，本章引入了扩散过程中生成的像素级注意力图 $A_i^{s,p}$ 。

例如在左上方的注意力图中，显著性检测任务需要识别图像中最显著的主体（即人物）。与此同时，边界检测任务的特征需要明确区分边界内部与外部区域。因此，显著性检测任务会优先关注人体区域的特征分布。

左下方的注意力图显示，人体解析任务借用了语义分割背景区域的特征。由于人体解析要求不同身体部位（如头部、手臂）的特征具有相似性，自然会对人物主体区域的特征分配较低关注度，注意力图也印证了这一点。当涉及角落处的高注意力值时，首先需要深入分析推理过程的细节。在推理过程中，本章将图像填充至 512×512 分辨率，此时注意力图对填充区域（即图像的实际背景）赋予较高关注值。在可视化过程中，为保持图像清晰度对填充区域进行了裁剪，但裁剪后的图像仍残留部分高关注值区域，即图中角落处的高注意力值。

右上方的注意力图聚焦于地面阴影区域与左侧长椅轮廓，这些区域的边界标注正是边界检测的重点关注区域。语义分割任务可利用边界检测分支提取的边界内部特征，有效区分背景区域与人物、犬类主体区域。右下方的注意力图则集中于人物手部轮廓，该区域同时是人体解析任务的关键分割目标之一。

单任务学习有效性分析：作为密集预测方法，TaskDiffusion 经过适当调整也可应用于单任务学习场景。当任务数量缩减为单一任务时，由于仅存在单个任务，跨任务关系图将失去作用。移除该模块后，本方法即可执行单任务学习。本章在单任务场景下进行实验，结果如表 4.7所示。相较于单任务学习基线，性能提升并没有多任务学习场景下显著。由于本章提出的跨任务编码机制旨在解决多任务学习的性能瓶颈，本方法通过扩散过程中捕获不同任务间的关系图，在多任务学习场景中具有更显著的效能优势。

初始种子的稳健性分析： 由于本方法基于扩散模型，其推理过程会受到噪声初始化的影响。为评估初始化的影响，本章在 表 4.8中列示了相同训练模型在不同初始种子下生成的三个结果。出于效率考量，采用损失值作为边界检测性能的评估指标。可见性能差异小于 0.01%，该误差范围可忽略不计。这表明本

表 4.7 单任务学习场景下的有效性消融实验。

设置	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	Δ_m % ↑
基线网络 (单任务学习场景)	79.18	69.57	85.28	13.45	75.60	-
本章方法 (单任务学习场景)	79.54	70.71	85.35	13.37	77.19	0.97
基线网络	77.34	66.17	85.18	13.78	72.40	-2.80
本章方法	78.83	67.40	85.31	13.38	74.68	-0.84

方法对不同噪声初始化具有强鲁棒性。

表 4.8 初始种子稳健性的消融实验。

设置	Semseg ↑	Parsing↑	Sal.↑	Nor. ↓	Bound. loss ↓
种子 1	81.2058	69.6165	84.9360	13.5463	0.04273246
种子 2	81.2060	69.6164	84.9360	13.5463	0.04273244
种子 3	81.2058	69.6165	84.9361	13.5463	0.04273246

第四节 本章小结

本章提出了一种新颖的基于扩散模型的多任务密集预测方法 TaskDiffusion。为使扩散模型适配多任务密集预测，本章设计了联合去噪扩散过程。首先，将任务特定标签编码至任务联合特征空间。这种统一编码策略避免了繁琐的任务特定编码，同时在标签编码中捕获任务间关联。进一步地，本章提出基于任务特定多层级特征的跨任务扩散解码器。该方法在保持高效性的同时，显式建模不同任务与层级间的交互关系。实验结果表明，本方法在所有任务上的性能均显著超越现有方法。

第五章 总结展望

多任务密集预测是计算机视觉中的一个重要的任务，能够助力神经网络在边缘设备上的高效部署。为了改进现有的多任务密集预测方法，本文从模型结构设计和建模范式两个方面深入探讨了现有方法的局限，提出了改进的思路和方法，并且用实验验证和分析了方法的有效性。

在模型结构方面，本文基于前沿的动态神经网络结构：混合专家模型展开探索，指出现有的混合专家模型存在两方面局限。第一，混合专家模型因为需要设置多个专家网络，所以在参数量和计算成本上对多任务模型产生较大负担。现有的方法为了保证效率往往会减少混合专家模型的专家数量，限制了模型潜力的充分发挥。第二，混合专家模型的路由算法难以建立全局任务关系，而这对于多任务密集预测模型的性能来说十分关键。针对第一方面局限，本文从低秩适应上获得灵感，认为负责不同任务的模型之间在参数上的差异可以用低秩矩阵表现。因此，本文引入低秩结构，对不同专家网络的矩阵参数施加显式的低秩约束，并利用矩阵低秩分解的方式进行轻量化。针对第二方面局限，本文引入一条平行的任务共享支路，由所有任务的梯度进行更新，从而建立在所有任务之间的关系。在两个基准数据集和总计六个任务上的实验表明，本文所提出方法在所有任务上的性能均超越了现有最优方法，同时效率也总体优于现有方法，广泛的消融实验也证明了各模块的有效性。

在建模范式方面，本文探索生成式方法在多任务密集预测中的应用，基于目前最为前沿的生成式方法扩散模型进行研究，揭示了其在多任务密集预测上的局限。首先，扩散模型的多步迭代去噪会因为任务数量的提升而放大其在效率上的负担。其次，现有方法仅仅研究扩散模型在捕获单一任务掩码内部分布的能力，忽视了捕获任务间关系。针对这两点局限，本文提出了创新的联合去噪扩散过程，将多个任务的去噪过程合并在一起，从而缓解了多步去噪带来的效率负担。此外，得益于多任务的联合去噪，本文还在去噪过程中显式地建模任务与任务之间的关系，提升了任务总体的性能。最后，本文在两个基准数据集上测试了本文所提出方法的性能，发现其在所有任务上都能带来明显的性能提升，与最先进方法结合之后，性能能超越现有的所有方法。同时，广泛的消融

实验与可视化结果也可以为本文的贡献提供更加深入的理解。

最后，本文阐述基于低秩混合专家模型和扩散模型的两个算法中待研究的问题，以及对其未来发展的展望。对于低秩混合专家模型来说，本方法虽然考虑了全标注的多任务密集预测场景，但是在实际应用中，训练的数据集往往是只有部分标注的，因此如何利用混合专家模型的灵活性并将其适配到部分标注场景上是一个非常值得研究的问题。对于扩散模型来说，首先，本文在 4-5 个任务组合场景中进行测试，并证明了低秩混合专家模型的优越性能，但任务更多时模型效果仍需深入研究。其次，本方法聚焦于具有相似性的任务组合，对差异显著的任务组合（例如实例分割与图像分类）探索也具有重要的研究意义。总而言之，希望本文工作能够为未来的研究人员提供启发，推动多任务密集预测这一领域的继续发展。

参考文献

[1] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 4700–4708.

[2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 3431–3440.

[3] 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 Trans. Pattern Anal. Mach. Intell., 2017, 40 (4): 834–848.

[4] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 1925–1934.

[5] YU C, WANG J, PENG C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation. [C] // Eur. Conf. Comput. Vis. 2018: 325–341.

[6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2016, 39 (6): 1137–1149.

[7] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction. [C] // Int. Conf. Comput. Vis. 2021: 12179–12188.

[8] BHAT S F, ALHASHIM I, WONKA P. Adabins: Depth estimation using adaptive bins. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 4009–4018.

[9] ISHIHARA K, KANERVISTO A, MIURA J, et al. Multi-task learning with attention for end-to-end autonomous driving. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 2902–2911.

[10] YU F, CHEN H, WANG X, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 2636–2645.

[11] 王越, 曹家乐. 基于任务特征解耦的自动驾驶视觉联合感知模型. [J]. 激光与光电子学进展, 2024, 61 (22): 2215007.

[12] 阳运秋, 胡庆茂, 汪震, 等. 基于多任务学习血管分割的拟定量侧支评分在 sCTA 评估急性缺血性卒中侧支循环中的应用. [J]. 西安交通大学学报（医学版）, 2024, 45 (3): 497–507.

[13] MISRA I, SHRIVASTAVA A, GUPTA A, et al. Cross-stitch networks for multi-task learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 3994–4003.

[14] GAO Y, MA J, ZHAO M, et al. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 3205–3214.

[15] YE H, XU D. TaskExpert: Dynamically Assembling Multi-Task Representations with Memorial Mixture-of-Experts. [C] // Int. Conf. Comput. Vis. 2023.

[16] VANDENHENDE S, GEORGOULIS S, VAN GANSBEKE W, et al. Multi-task learning for dense prediction tasks: A survey. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2021, 44 (7): 3614–3633.

[17] VANDENHENDE S, GEORGOULIS S, VAN GOOL L. Mti-net: Multi-scale task interaction networks for multi-task learning. [C] // Eur. Conf. Comput. Vis. Springer. 2020: 527–543.

[18] 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 Conf. Comput. Vis. Pattern Recog. 2018: 675–684.

[19] YE H, XU D. Inverted pyramid multi-task transformer for dense scene understanding. [C] // Eur. Conf. Comput. Vis. Springer. 2022: 514–530.

[20] ZHANG Z, CUI Z, XU C, et al. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 4106–4115.

[21] LE M.-Q, NGUYEN T V, LE T.-N, et al. Maskdiff: Modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation. [C] // Proceedings of the AAAI conference on artificial intelligence. Vol. 38. 3. 2024: 2874–2881.

[22] CHEN T, SAXENA S, LI L, et al. Pix2seq: A language modeling framework for object detection. [J]. ArXiv preprint arXiv:2109.10852, 2021.

[23] WANG H, CAO J, ANWER R M, et al. DFormer: Diffusion-guided Transformer for Universal Image Segmentation. [J]. ArXiv preprint arXiv:2306.03437, 2023.

[24] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics. [C] // Int. Conf. Mach. Learn. PMLR. 2015: 2256–2265.

[25] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models. [J]. Adv. Neural Inform. Process. Syst., 2020, 33: 6840–6851.

[26] HO J, CHAN W, SAHARIA C, et al. Imagen video: High definition video generation with diffusion models. [J]. ArXiv preprint arXiv:2210.02303, 2022.

[27] JI Y, CHEN Z, XIE E, et al. Ddp: Diffusion model for dense visual prediction. [J]. ArXiv preprint arXiv:2303.17559, 2023.

[28] SONG J, MENG C, ERMON S. Denoising diffusion implicit models. [J]. ArXiv preprint arXiv:2010.02502, 2020.

[29] ZHOU L, CUI Z, XU C, et al. Pattern-structure diffusion for multi-task learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 4514–4523.

[30] BRÜGGEMANN D, KANAKIS M, OBUKHOV A, et al. Exploring relational context for multi-task dense prediction. [C] // Int. Conf. Comput. Vis. 2021: 15869–15878.

[31] VANDENHENDE S, GEORGOULIS S, DE BRABANDERE B, et al. Branched multi-task networks: deciding what layers to share. [C] // Brit. Mach. Vis. Conf. 2019.

[32] BRUGGEMANN D, KANAKIS M, GEORGOULIS S, et al. Automated search for resource-efficient branched multi-task networks. [J]. ArXiv preprint arXiv:2008.10292, 2020.

[33] FAN Z, SARKAR R, JIANG Z, et al. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. [C] // Adv. Neural Inform. Process. Syst. Vol. 35. 2022: 28441–28457.

[34] CHEN Z, SHEN Y, DING M, et al. Mod-Squad: Designing Mixtures of Experts As Modular Multi-Task Learners. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2023: 11828–11837.

[35] CHEN T, CHEN X, DU X, et al. AdaMV-MoE: Adaptive Multi-Task Vision Mixture-of-Experts. [C] // Int. Conf. Comput. Vis. 2023: 17346–17357.

[36] JACOBS R A, JORDAN M I. Learning piecewise control strategies in a modular neural network architecture. [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1993, 23 (2): 337–345.

[37] JACOBS R A, JORDAN M I, NOWLAN S J, et al. Adaptive mixtures of local experts. [J]. Neural computation, 1991, 3 (1): 79–87.

[38] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 7482–7491.

[39] CHEN Z, BADRINARAYANAN V, LEE C.-Y, et al. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. [C] // Int. Conf. Mach. Learn. PMLR. 2018: 794–803.

[40] GUO M, HAQUE A, HUANG D.-A, et al. Dynamic task prioritization for multitask learning. [C] // Eur. Conf. Comput. Vis. 2018: 270–287.

[41] ZHAO X, LI H, SHEN X, et al. A modulation module for multi-task learning with applications in image retrieval. [C] // Eur. Conf. Comput. Vis. 2018: 401–416.

[42] CHEN Z, NGIAM J, HUANG Y, et al. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. [J]. Adv. Neural Inform. Process. Syst., 2020, 33: 2039–2050.

[43] CHEN T, LI L, SAXENA S, et al. A generalist framework for panoptic segmentation of images and videos. [C] // Int. Conf. Comput. Vis. 2023: 909–919.

[44] SAXENA S, KAR A, NOROUZI M, et al. Monocular depth estimation using diffusion models. [J]. ArXiv preprint arXiv:2302.14816, 2023.

[45] LEE H.-Y, TSENG H.-Y, YANG M.-H. Exploiting Diffusion Prior for Generalizable Dense Prediction. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2024: 7861–7871.

[46] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 2881–2890.

[47] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation. [C] // Int. Conf. Comput. Vis. 2019: 9167–9176.

[48] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation. [C] // Int. Conf. Comput. Vis. 2019: 603–612.

[49] PENG C, ZHANG X, YU G, et al. Large kernel matters–improve semantic segmentation by global convolutional network. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 4353–4361.

[50] CHEN L.-C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation. [J]. ArXiv preprint arXiv:1706.05587, 2017.

[51] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation. [C] // Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer. 2015: 234–241.

[52] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. [J]. Adv. Neural Inform. Process. Syst., 2021, 34: 12077–12090.

[53] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 6881–6890.

[54] CHENG B, SCHWING A, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation. [J]. Adv. Neural Inform. Process. Syst., 2021, 34: 17864–17875.

[55] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2022: 1290–1299.

[56] 孙博远, 刘夏雷, 侯淇彬. 基于深度学习的半监督语义分割综述. [J]. 北京交通大学学报, 2024 (5).

[57] 计梦予, 裘肖明, 于治楼. 基于深度学习的语义分割方法综述. [J]. 信息技术与信息化, 2017 (10): 4.

[58] 陈鸿翔. 基于卷积神经网络的图像语义分割. [D]. 浙江大学, 2016.

[59] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. [C] // Int. Conf. Comput. Vis. 2015: 2650–2658.

[60] FACIL J M, UMMENHOFER B, ZHOU H, et al. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 11826–11835.

[61] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 270–279.

[62] LAINA I, RUPPRECHT C, BELAGIANNIS V, et al. Deeper depth prediction with fully convolutional residual networks. [C] // 2016 Fourth international conference on 3D vision (3DV). IEEE. 2016: 239–248.

[63] LIU F, SHEN C, LIN G. Deep convolutional neural fields for depth estimation from a single image. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 5162–5170.

[64] 江俊君, 李震宇, 刘贤明. 基于深度学习的单目深度估计方法综述. [J]. 计算机学报, 2022 (006): 045.

[65] LI G, YU Y. Deep contrast learning for salient object detection. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 478–487.

[66] KUEN J, WANG Z, WANG G. Recurrent attentional networks for saliency detection. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 3668–3677.

[67] LIU N, HAN J. Dhsnet: Deep hierarchical saliency network for salient object detection. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 678–686.

[68] 王自全, 张永生, 于英, 等. 深度学习背景下视觉显著性物体检测综述. [J]. 中国图象图形学报, 2022, 27 (7): 17.

[69] GALLIANI S, SCHINDLER K. Just look at the image: viewpoint-specific surface normal prediction for improved multi-view reconstruction. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 5479–5487.

[70] CANNY J. A computational approach to edge detection. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 1986 (6): 679–698.

[71] ZHENG Z, ZHA B, YUAN H, et al. Adaptive edge detection algorithm based on improved grey prediction model. [J]. IEEE Access, 2020, 8: 102165–102176.

[72] LI B, SHEN C, DAI Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 1119–1127.

[73] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. [C] // Int. Conf. Learn. Represent. 2020.

[74] RAVI N, GABEUR V, HU Y.-T, et al. Sam 2: Segment anything in images and videos. [J]. ArXiv preprint arXiv:2408.00714, 2024.

[75] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything. [C] // Int. Conf. Comput. Vis. 2023: 4015–4026.

[76] YANG L, KANG B, HUANG Z, et al. Depth anything: Unleashing the power of large-scale unlabeled data. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2024: 10371–10381.

[77] RUDER S, BINGEL J, AUGENSTEIN I, et al. Latent multi-task architecture learning. [C] // Proceedings of the AAAI conference on artificial intelligence. Vol. 33. 01. 2019: 4822–4829.

[78] LU Y, KUMAR A, ZHAI S, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 5334–5343.

[79] GUO P, LEE C.-Y, ULBRICHT D. Learning to branch for multi-task learning. [C] // Int. Conf. Mach. Learn. PMLR. 2020: 3854–3863.

[80] LIU S, JOHNS E, DAVISON A J. End-to-end multi-task learning with attention. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 1871–1880.

[81] YE H, XU D. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. [C] // Int. Conf. Learn. Represent. 2022.

[82] ZHANG Z, CUI Z, XU C, et al. Joint task-recursive learning for semantic segmentation and depth estimation. [C] // Eur. Conf. Comput. Vis. 2018: 235–251.

[83] HUANG S, LI X, CHENG Z.-Q, et al. Gnas: A greedy neural architecture search method for multi-attribute learning. [C] // ACM Int. Conf. Multimedia. 2018: 2049–2057.

[84] OQUAB M, DARCET T, MOUTAKANNI T, et al. Dinov2: Learning robust visual features without supervision. [J]. ArXiv preprint arXiv:2304.07193, 2023.

[85] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. [C] // Int. Conf. Comput. Vis. 2021: 10012–10022.

[86] QU X, DONG D, HU X, et al. Llama-moe v2: Exploring sparsity of llama from perspective of mixture-of-experts with post-training. [J]. ArXiv preprint arXiv:2411.15708, 2024.

[87] ZHU T, QU X, DONG D, et al. Llama-moe: Building mixture-of-experts from llama with continual pre-training. [C] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 15913–15923.

[88] CLARK A, DE LAS CASAS D, GUY A, et al. Unified scaling laws for routed language models. [C] // Int. Conf. Mach. Learn. PMLR. 2022: 4057–4086.

[89] LI L, WU Z, JI Y. Mote: Mixture of Task-Specific Experts for Pre-Trained Model-Based Continual Learning. [J]. Available at SSRN 5035279,

[90] LE M, NGUYEN H, NGUYEN T, et al. Mixture of experts meets prompt-based continual learning. [J]. Adv. Neural Inform. Process. Syst., 2024, 37: 119025–119062.

[91] ZHANG D, ZHANG K, CHU S, et al. MORE: A MIXTURE OF LOW-RANK EXPERTS FOR ADAPTIVE MULTI-TASK LEARNING. [J].

[92] UDELL M, HORN C, ZADEH R, et al. Generalized low rank models. [J]. Foundations and Trends® in Machine Learning, 2016, 9 (1): 1–118.

[93] IDELBAYEV Y, CARREIRA-PERPINÁN M A. Low-rank compression of neural nets: Learning the rank of each layer. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 8049–8059.

[94] YANG Y, HOSPEDALES T M. Trace norm regularised deep multi-task learning. [J]. ArXiv preprint arXiv:1606.04038, 2016.

[95] SU C, YANG F, ZHANG S, et al. Multi-task learning with low rank attribute embedding for person re-identification. [C] // Int. Conf. Comput. Vis. 2015: 3739–3747.

[96] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models. [J]. ArXiv preprint arXiv:2106.09685, 2021.

[97] LIU Y.-C, MA C.-Y, TIAN J, et al. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. [J]. Adv. Neural Inform. Process. Syst., 2022, 35: 36889–36901.

[98] JIE S, DENG Z.-H. Fact: Factor-tuning for lightweight adaptation on vision transformer. [C] // Proceedings of the AAAI conference on artificial intelligence. Vol. 37. 1. 2023: 1060–1068.

[99] SUNG Y.-L, CHO J, BANSAL M. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2022: 5227–5237.

[100] AGHAJANYAN A, ZETTLEMOYER L, GUPTA S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. [J]. ArXiv preprint arXiv:2012.13255, 2020.

[101] KOLESNIKOV A, BEYER L, ZHAI X, et al. Big transfer (bit): General visual representation learning. [C] // Eur. Conf. Comput. Vis. Springer. 2020: 491–507.

[102] LI X, THICKSTUN J, GULRAJANI I, et al. Diffusion-lm improves controllable text generation. [J]. Adv. Neural Inform. Process. Syst., 2022, 35: 4328–4343.

[103] CHEN T, ZHANG R, HINTON G. Analog bits: Generating discrete data using diffusion models with self-conditioning. [J]. ArXiv preprint arXiv:2208.04202, 2022.

[104] DIELEMAN S, SARTRAN L, ROSHANNAI A, et al. Continuous diffusion for categorical data. [J]. ArXiv preprint arXiv:2211.15089, 2022.

[105] CHEN C, DING H, SISMAN B, et al. Diffusion Models for Multi-Task Generative Modeling. [C] // Int. Conf. Learn. Represent. 2023.

[106] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2022: 10684–10695.

[107] KE B, OBUKHOV A, HUANG S, et al. Repurposing diffusion-based image generators for monocular depth estimation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2024: 9492–9502.

[108] ZHU M, LIU Y, LUO Z, et al. Unleashing the potential of the diffusion model in few-shot semantic segmentation. [J]. ArXiv preprint arXiv:2410.02369, 2024.

[109] KINGMA D P, WELLING M. Auto-encoding variational bayes. [J]. ArXiv preprint arXiv:1312.6114, 2013.

[110] YE H, XU D. DiffusionMTL: Learning Multi-Task Denoising Diffusion Model from Partially Annotated Data. [J]. ArXiv preprint arXiv:2403.15389, 2024.

[111] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. [J]. Int. Conf. Mach. Learn., 2022, 23 (1): 5232–5270.

[112] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 13713–13722.

[113] HOU Q, ZHANG L, CHENG M.-M, et al. Strip pooling: Rethinking spatial pooling for scene parsing. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 4003–4012.

[114] DING X, ZHANG X, HAN J, et al. Diverse branch block: Building a convolution as an inception-like unit. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 10886–10895.

[115] DING X, ZHANG X, MA N, et al. Repvgg: Making vgg-style convnets great again. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 13733–13742.

[116] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. [J]. ArXiv preprint arXiv:1701.06538, 2017.

[117] RIQUELME C, PUIGCERVER J, MUSTAFA B, et al. Scaling vision with sparse mixture of experts. [J]. Adv. Neural Inform. Process. Syst., 2021, 34: 8583–8595.

[118] CHEN X, MOTTAGHI R, LIU X, et al. Detect what you can: Detecting and representing objects using holistic models and body parts. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2014: 1971–1978.

[119] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from rgbd images. [C] // Eur. Conf. Comput. Vis. Springer. 2012: 746–760.

[120] MANINIS K.-K, RADOSAVOVIC I, KOKKINOS I. Attentive single-tasking of multiple tasks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 1851–1860.

[121] CHEN T, SAXENA S, LI L, et al. A unified sequence interface for vision tasks. [J]. Adv. Neural Inform. Process. Syst., 2022, 35: 31333–31346.

[122] NICHOL A Q, DHARIWAL P. Improved denoising diffusion probabilistic models. [C] // Int. Conf. Mach. Learn. PMLR. 2021: 8162–8171.

[123] XU Y, LI X, YUAN H, et al. Multi-task learning with multi-query transformer for dense prediction. [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[124] XU Y, YANG Y, ZHANG L. DeMT: Deformable mixer transformer for multi-task learning of dense prediction. [C] // Proceedings of the AAAI conference on artificial intelligence. Vol. 37. 3. 2023: 3072–3080.

[125] WANG S, LI J, ZHAO Z, et al. Tsp-transformer: Task-specific prompts boosted transformer for holistic scene understanding. [C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 925–934.

[126] YANG Y, JIANG P.-T, HOU Q, et al. Multi-Task Dense Prediction via Mixture of Low-Rank Experts. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2024: 27927–27937.

[127] YANG Y, JIANG P.-T, HOU Q, et al. Multi-Task Dense Prediction via Mixture of Low-Rank Experts. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2024.

[128] PARK B, WOO S, GO H, et al. Denoising task routing for diffusion models. [J]. ArXiv preprint arXiv:2310.07138, 2023.

致谢

时光飞逝，研究生阶段的求学生活转眼就要迎来尾声，立于当下回首过去三年的时光，诸多感慨难以一一言明，但是对于研究之路上与我同行的导师与同学的感激之情却是无比鲜明的。

在此，我想要对我的导师程明明老师表达感谢，自把我领进科研的世界开始，程老师就在各个方面给予我诸多指导，不论在学术研究上还是为人处世上都让我受益匪浅。此外，我还要感谢侯淇彬老师，侯老师的帮助让我逐渐找到科研的窍门。最后，我也要感谢实验室的同窗伙伴，在积极向上的学术氛围里，我们共同为了科研而努力，在迷茫的时候相互讨论，在沮丧的时候相互鼓励，在实验成功的时候相互发自内心地祝贺。在今后的日子里，我相信这一段求学的经历会成为人生中难以忘怀的一笔。

个人简历

杨雨奇，出生于 2000 年 6 月 14 日。在 2022 年毕业于南开大学计算机科学与技术专业并获得学士学位。于 2022 年至今在南开大学就读计算机科学与技术研究生。

研究生期间发表论文:

- ***Multi-Task Dense Prediction via Mixture of Low-Rank Experts***
  Yuqi Yang*, Peng-Tao Jiang*, Qibin Hou, Hao Zhang, Jinwei Chen, Bo Li. IEEE Computer Vision and Pattern Recognition (CVPR) 2024
- ***Traffic Scene Parsing through the TSP6K Dataset***
  Peng-Tao Jiang*, **Yuqi Yang***, Yang Cao, Qibin Hou, Ming-Ming Cheng, Chunhua Shen. IEEE Computer Vision and Pattern Recognition (CVPR) 2024
- ***CorrMatch: Label Propagation via Correlation Matching for Semi-Supervised Semantic Segmentation***
  Boyuan Sun, **Yuqi Yang**, Weifeng Yuan, Le Zhang, Ming-Ming Cheng, Qibin Hou. IEEE Computer Vision and Pattern Recognition (CVPR) 2024
- ***Multi-Task Dense Predictions via Unleashing the Power of Diffusion.***
  Yuqi Yang*, Peng-Tao Jiang*, Qibin Hou, Hao Zhang, Jinwei Chen, Bo Li. International Conference on Learning Representations. (ICLR) 2025