北京交通大学

本科毕业设计（论文）

联合单目深度估计的深度图像超分辨率重建算法研究

Research on Joint Depth Map Super-Resolution and Monocular Depth Estimation Algorithm

学　　院：软件学院

专　　业：软件工程

学生姓名：唐麒

学　　号：17301138

指导教师：冯凤娟

北京交通大学

2022 年 5 月

学士论文版权使用授权书

本学士论文作者完全了解北京交通大学有关保留、使用学士论文的规定。特授权北京交通大学可以将学士论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：  　　指导教师签名： 

签字日期：2021 年 6 月 11 日　　签字日期：2021 年 6 月 11 日

中文摘要

摘要：深度图像超分辨率重建是工业中具有较高实际应用需求的任务。现有的颜色引导的深度图像超分辨率重建方法通常需要额外的分支来从彩色图像中提取高频信息，用于指导低分辨率深度图像的重建。但是，由于彩色图像和深度图像两种模态之间仍然存在着一些差异，因此在特征维度或边缘图维度上进行直接的信息传输无法获得令人满意的结果，甚至可能在 RGB-D 图像对之间结构不一致的区域造成纹理复制等问题。

受多任务学习的启发，本文提出了联合单目深度估计的深度图像超分辨率重建网络。由于两个任务可以共享训练数据，因而在将两个任务统一于一个联合学习网络时，无需引入其他监督标签。对于两个子网的交互，本文采用了差异化的指导策略，并相应地设计了两个桥接器。一个是在特征编码过程中的高频注意力桥，它被设计用于从单目深度估计任务中学习高频信息以指导深度图像超分辨率重建任务。另一个是在深度图像重建过程中的内容引导桥，它被设计用于为单目深度估计任务提供从深度图像超分辨率重建任务中学到的内容指导。整个网络体系结构具有高度的可移植性，可以为关联深度图像超分辨率重建任务和单目深度估计任务提供范例。在公开的基准数据集上进行的实验表明，本文提出的方法取得了具有竞争优势的性能。

高质量和高分辨率的深度图像在自动驾驶、三维重建、人机交互和虚拟现实等诸多领域至关重要，因而更好地从低分辨率的深度图像恢复出高分辨率的深度图像将有助于推动下游任务的发展和实际应用。

关键词：深度图像；超分辨率重建；单目深度估计；多任务学习

ABSTRACT

ABSTRACT: Depth map super-resolution is a task with high practical application requirements in the industry. Existing color-guided depth map super-resolution methods usually necessitate an extra branch to extract high-frequency detail information from RGB image to guide the low-resolution depth map reconstruction. However, because there are still some differences between the two modalities, direct information transmission in the feature dimension or edge map dimension cannot achieve satisfactory result, and may even trigger texture copying in areas where the structures of the RGB-D pair are inconsistent.

Inspired by the multi-task learning, we propose a joint learning network of depth map super-resolution (DSR) and monocular depth estimation (MDE) without introducing additional supervision labels. For the interaction of two subnetworks, we adopt a differentiated guidance strategy and design two bridges correspondingly. One is the high-frequency attention bridge (HABdg) designed for the feature encoding process, which learns the high-frequency information of the MDE task to guide the DSR task. The other is the content guidance bridge (CGBdg) designed for the depth map reconstruction process, which provides the content guidance learned from DSR task for MDE task. The entire network architecture is highly portable and can provide a paradigm for associating the DSR and MDE tasks. Extensive experiments on benchmark datasets demonstrate that our method achieves competitive performance.

In many fields, such as autonomous navigation, 3D reconstruction, human-computer interaction, and virtual reality, a high-quality and high-resolution depth map is needed. Therefore, improving the reconstruction of high-resolution depth map from low-resolution depth map will promote the development and practical application of downstream tasks.

KEYWORDS： Depth map; Super-resolution; Monocular Depth Estimation; Multi-task Learning

目　　录

1 引言

本章将主要介绍深度图像超分辨率重建的研究背景、研究现状、研究难点以及本文的研究内容与主要贡献。本章的组织结构如下：第一部分主要介绍深度图像超分辨率重建的研究背景以及该任务在实际应用中的研究意义和价值，第二部分将介绍深度图像超分辨率重建和单目深度估计的研究现状以及将两个任务统一于一个网络框架联合学习的主要难点，第三部分将介绍本文的主要研究内容和贡献，最后介绍论文的结构安排。

1.1 研究背景与实际应用

在理解场景时，人们不仅可以感知其外观（例如颜色，纹理等），还可以捕获深度信息以产生立体感。更好的场景理解可以促进自动驾驶[1]，三维重建[2]等依赖于高质量和高分辨率深度信息领域的研究。便携式消费级深度相机（如 Microsoft Kinect 和 Lidar）的出现和普及，为准确快速地获取场景深度提供了极大的便利。但是，由于当前深度相机成像能力的限制，深度图像的分辨率通常较低，无法与同场景的高分辨率彩色图像相匹配。面对诸多应用领域对高质量深度图像的需求[3-5]，深度图像超分辨率重建技术作为解决方案获得了越来越多的关注。

深度图像超分辨率重建技术是指在不改变深度相机或深度传感器的前提下，通过算法恢复出相机或传感器截止频率以外的高频信息（图像的高频信息是指灰度变化快速的区域，例如物体边缘），同时改善成像时的模糊现象并有效抑制图像中的随机噪声，从而重建出高质量、高分辨率的深度图像。传统的基于滤波的方法[6-7]和基于优化的方法[8-9]使用人工构造的滤波器或目标函数很难恢复出准确的高分辨率深度图像。近年来，随着深度学习的快速发展，许多研究已经表明了其在深度图像超分辨率重建任务上的有效性[10-11]。基于卷积神经网络的深度图像超分辨率重建技术可以自动地从数据中学习更强的特征表示来重建高分辨率深度图像。

在实际应用中，高分辨率的彩色图像易于获得，且与深度图像具有很强的结构相似性，因而可以为深度图像超分辨率重建提供一些先验信息。这种利用彩色图像引导的深度图像超分辨率重建算法被称为颜色（或纹理图像）指导的深度图像超分辨率重建算法。现有的颜色指导的深度图像超分辨率重建算法通常需要一个额外的分支来从彩色图像中获取丰富的指导信息[12]，然后利用它们来指导超分辨率重建分支对深度图像的特征提取。但是，彩色图像的结构并不总是与深度图像相一致。如图 1-1 所示，绿色矩形区域在彩色图像上具有复杂的纹理变化，但是由于目标内部的深度一致性，在对应的深度图像中该区域并没有表现出对应的纹理结构。因此，如果仅将从彩色图像提取的特征或边

缘特征传递到超分辨率重建分支，就很容易造成诸如纹理复制和深度流失等问题。因此，探索如何有效地提取和利用高分辨率彩色图像蕴含的信息对于深度图像超分辨率重建任务非常重要。

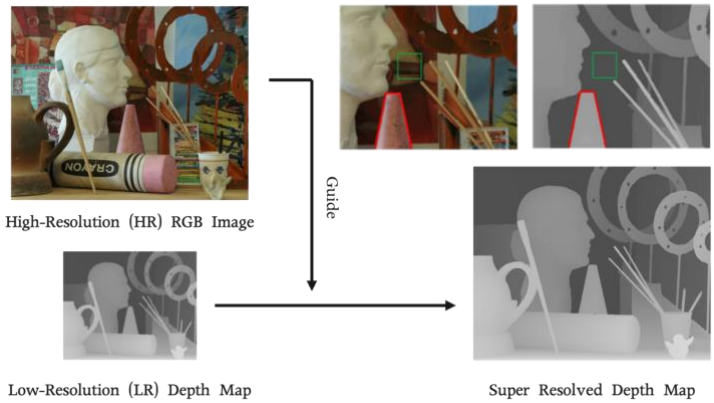

图 1-1 颜色引导的深度图像超分辨率重建示意图

为了寻找上述问题的解决方案，本文将目光聚焦在了另一个与深度图像有关的任务——单目深度估计。单目深度估计旨在将场景从光度表示映射到几何表示。具体而言，单目深度估计的输入是一幅彩色图像，而输出则是估计的深度图像。

单目深度估计与深度图像超分辨率重建两个任务是天然相关的：

（1） 将两个任务嵌入联合学习的训练框架中，无需引入额外的监督标签（例如语义标签），即两个任务的训练数据集可以共享。

（2） 由于单目深度估计可以在连续的训练和学习过程中实现从彩色图像到深度图像的跨模态信息转换，因而面向单目深度估计学习到的彩色图像特征更适合指导深度图像超分辨率重建。

综上所述，深度图像超分辨率重建与单目深度估计的联合学习可以在不增加监督信息的情况下实现更好的颜色指导。

1.2 研究现状及难点

本部分将分别介绍深度图像超分辨率重建和单目深度估计的研究现状，并将详细分析将单目深度估计与深度图像超分辨率重建统一于联合学习框架的难点。

1.2.1 深度图像超分辨率重建

由于彩色图像和深度图像间的结构相似性，现有的许多方法都利用颜色信息指导低分辨率深度图像的重建。Zhao 等[13]提出提出了彩色-深度条件生成对抗网络（Color-Depth Conditional Generative Adversarial Network, CDcGAN）来同时解决 3D 视频中的深

度图像和彩色图像超分辨率重建问题，该网络考虑了同一场景下彩色图像和深度图像的结构相似性，采用彩色图像和深度图像的交互信息来相互促进彼此的性能。 Hui 等[10]设计了多尺度引导的卷积神经网络（Multi-Scale Guided convolutional network, MSG），将从彩色图像中提取的丰富层次特征用于改善深度图像超分辨率重建过程中图像的模糊现象。Zuo 等[14]提出了彩色引导的深度图像增强残差稠密网络（Residual Dense Network for Guided Depth Enhancement, RDN-GDE），将多尺度残差与稠密连接相结合以逐步增强低分辨率的深度图像。 Ye 等[15]提出了渐进的多分支聚合网络（Progressive Multi-Branch Aggregation network, PMBA），通过重建分支和引导分支融合的方式逐步优化反卷积得到的高分辨率深度图像。 Guo 等[16]提出了层次特征驱动的深度图像超分辨率重建残差网络 DepthSR-Net, 利用 U-Net 结构对插值后的深度图像进行编码，并在解码过程中与相应尺度的彩色特征进行融合。

1.2.2 单目深度估计

单目深度估计是一个典型的不适定逆问题，这是由于其将在信息不足以完全指定解决方案的情况下尝试恢复一些未知数。与基于左右视图的深度估计任务相比，单目深度估计具有更加广阔的实际应用前景，但目前单目深度估计的性能仍然非常有限。Eigen 等[17]设计了一个包含对场景全局的粗估计和局部区域的精估计两个尺度的卷积神经网络，开创了深度学习在单目深度估计领域的先河。Laina 等[18]使用了更深的残差网络并通过小卷积进行上采样，从而提升了单目深度估计的效率。Cao 等[19]将原始的连续深度离散为固定数量的深度范围，进而将单目深度估计从回归任务转换为分类任务，观察到了明显的性能提升。

1.2.3 面向深度图像的多任务联合学习

多任务学习的目的是通过相关任务联合以提升特定任务的性能，最简单的联合方法是通过损失函数将两个或多个任务组合在一起。但由于在网络设计中缺乏交互性，这通常不是最佳的策略。目前，许多与深度图像处理有关的任务都采用了多任务学习策略。Zhang 等[20]提出的联合任务递归学习（Joint Task-Recursive Learning, TRL）框架将问题序列化为任务交替的时间序列来递归地优化语义分割和单目深度估计任务的结果。这类方法仅仅通过网络和损失约束的设计来驱动两个任务的联合学习，而没有明确地探究任务之间存在的约束。He 等[21]基于两个任务的几何约束提出联合语义分割的单目深度估计网络（Semantic Object Segmentation and Depth Estimation Network, SOSD-Net）这一工作明确探究了多任务学习中任务之间的关联性，但单目深度估计和语义分割的相关性仍

然较弱。此外，相对于单目深度估计而言，这些方法在训练时需要引入额外的标签（如语义标签）。Sun 等[22]提出了一种通过单目深度估计在训练过程中帮助深度图像超分辨率重建更好地理解场景结构的知识蒸馏方法，从而证明了单目深度估计在改善深度图像超分辨率重建性能方面的有效性。但其通过比较训练过程中两个任务的平均像素误差来确定两个任务交互方向的方法，并没有明确地探究深度图像超分辨率重建与单目深度估计的相关性。

综上，联合单目深度估计的深度图像超分辨率重建的研究难点在于探究两个任务的关联性并设计合理的交互方式，尤其是探究如何将深度估计学习到的“知识”通过多任务学习的模式帮助低分辨率的深度图像进行超分辨率重建，从而达到更好的重建效果。

1.3 本文主要工作

受上述分析的驱动，本文提出了深度图像超分辨率重建和单目深度估计的联合学习网络（BridgeNet），关注于通过有效地桥接两个任务以实现更好的深度图像超分辨率重建。本文基于编码器-解码器（Encoder-Decoder）结构分别为单目深度估计和深度图像超分辨率重建任务设计了两个子网络，即深度图像超分辨率重建子网络（DSRNet）和单目深度估计子网络（MDENet），它们以多任务学习的模式协同工作。此外，本文还分别在编码器和解码器中设计了两个不同的桥接器（Bridge）以实现两个子网络的差异化引导。具体而言，单目深度估计子网络的编码器从彩色图像中学习面向深度图像的多级特征，这样的彩色特征更适用于指导深度图像的超分辨率重建。因此，本文提出了一个高频注意力桥（HABdg），将单目深度估计学习到的高频信息用于指导深度图像超分辨率重建。单目深度估计子网络和深度图像超分辨率重建子网络的特征解码器用于进一步提取面向任务的特征，以进行深度估计和深度图像超分辨率重建。由于单目深度估计存在的尺度模糊性，单目深度估计相较于深度图像的超分辨率重建更难达到良好的性能。因此，遵循简单任务指导困难任务的原则，本文提出了内容引导桥（CGBdg），旨在使深度图像超分辨率重建子网络在深度特征空间为单目深度估计子网络提供内容引导。除了在模型设计层面上关联这两个任务外，本文还在损失函数方面对其进行约束，以期两个子网络可以相互促进。

综上所述，本文主要贡献如下：

（1） 本文在联合学习网络中将深度图像超分辨率重建任务和单目深度估计任务相关联，以提升深度图像超分辨率重建的性能。本文提出的联合学习网络包括深度图像超分辨率重建子网络（DSRNet）和单目深度估计子网络（MDENet），以及两个用于联合学习的桥接器，即高频注意力桥（HABdg）和内容引导桥（CGBdg）。本文的整个网络结构具有高度的可移植性，可以

为关联深度图像超分辨率重建和单目深度估计任务提供范例。此外，与其他多任务学习不同，本文用于联合学习的两个任务不需要引入其他的监督信息。

（2） 特征编码阶段中的高频注意力桥（HABdg）将从单目深度估计子网络学习到的彩色高频信息传输到深度图像超分辨率重建子网络，从而可以提供更接近深度模态的颜色指导信息。遵循简单任务指导困难任务的原则，本文在特征解码阶段切换了两个任务的指导角色，并提出了内容引导桥（CGBdg），从而可以让深度图像超分辨率重建子网络在深度特征空间为单目深度估计子网络提供内容引导。

（3） 在不引入其他监督信息的情况下，本文的方法在多个公开基准数据集上均达到了具有竞争力的性能。

1.4 结构安排

本文一共分为五章，每一章节的内容安排如下：

第一章为引言，以深度图像超分辨率重建的研究背景和现实意义、颜色引导的深度图像超分辨率重建存在的问题、单目深度估计与深度图像超分辨率重建颜色分支的相似性为脉络层层递进，进而引出本文在多任务学习模式下将深度估计和深度图像超分辨率重建相关联的研究内容，接着分别介绍了两个任务的研究现状以及将他们统一于一个网络框架联合学习的关键难点，最后对论文的结构组织进行介绍。

第二章为相关技术和方法介绍，主要介绍了与本文研究相关的技术和方法，即联合单目深度估计的深度图像超分辨率重建网络的基石。

第三章为算法实现，详细地介绍了单目深度估计和深度图像超分辨率重建联合学习网络的架构设计，包括深度图像超分辨率重建子网络，单目深度估计子网络以及高频注意力桥和内容引导桥。

第四章为实验测试与分析，介绍了本文了采用的公开基准数据集和评估指标，然后对比分析了本文设计的网络在所选公开基准数据集上进行训练和测试的性能表现，以及与其他最新的深度图像超分辨率重建网络性能的对比。此外，通过消融实验验证了不同设计对网络性能的影响。

第五章为结论，对本文的工作进行总结，囊括了本文的主要贡献，并概述了未来工作的研究方向。

1.5 本章小结

本章层层递进地介绍了深度图像超分辨率重建的研究背景和意义，然后从颜色引导

的深度图像超分辨率重建存在的问题出发，阐明了联合单目深度估计与深度图像超分辨率重建联合学习算法的研究动机，进而分别介绍了两个任务的研究现状以及探究两个任务的关联性在是设计两个任务联合学习网络的重点和难点，然后对本文的主要贡献进行了详细介绍。

2 相关技术和方法

随着深度学习和神经网络技术的发展，计算机视觉相关任务的性能获得了显著提升。这些任务的发展不仅得益于主干网络（Backbone）令人瞩目的特征提取能力，也受益于多个领域的技术和方法。图像超分辨率重建任务作为计算机视觉领域的一个重要分支，随着相关理论和技术的不断丰富和发展，研究者已逐步认知到与该任务相关的关键问题和难点问题，并借鉴和采用了相应的技术和方法以提升图像超分辨率重建任务的性能。在本章中，将主要介绍计算机视觉任务的通用方法和与图像超分辨率重建任务密切相关的方法，包括：卷积神经网络[23]、编码器-解码器架构[24]、注意力机制[25]、多尺度理论[15]和像素重组[26]。

2.1 卷积神经网络

卷积神经网络（Convolutional Neural Network， CNN）是一种具有稀疏连接，权重共享等特点的深层神经网络，一般由若干卷积层（Convolutional Layer）、池化层（Pooling Layer）、激活函数（Activation Function）、批正则化层（Batch Normalization Layer）和全连接层（Full Connected Layer）交叉堆叠而成。网络通过对损失函数（Loss Function）的优化（Optimizer）实现参数的更新，从而获得最优的参数。

2.1.1 卷积

顾名思义，卷积神经网络的核心是卷积操作。卷积，可以看作是特征提取器，图像经过卷积运算后可以得到图像的特征映射，如图 2-1 所示。给定一个图像（或特征图）$X \in R^{W \times H}$ 和一个卷积核 $\omega \in R^{w \times h}$，其中 $W \times H$ 分别为图像的宽度和高度和，$w \times h$ 分别为卷积核的宽度和高度，且一般有 $w \ll W$，$h \ll H$。则卷积操作可定义为式 2-1。

$$y_{ij} = \sum_{u=1}^{w} \sum_{v=1}^{H} \omega_{uv} \, x_{i-u+1, j-v+1} \qquad (2-1)$$

式中，y_{ij} —— 卷积的输出；

ω_{uv} —— 卷积核的一个位置；

$x_{i+u-1, j+v-1}$ —— 图像中与卷积核位置相对应的像素。

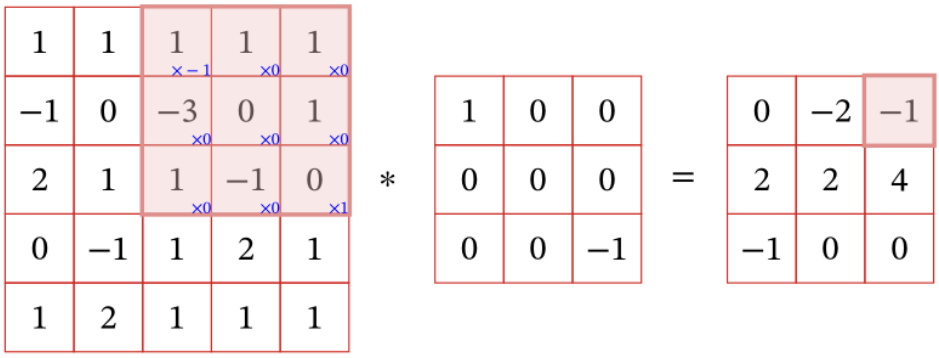

图 2-1 卷积示例图

如图所示，第一个矩阵为参与卷积操作的图像（或特征图），第二个矩阵为大小为 3×3 的卷积核，第三个矩阵则为将卷积核作用于图像（或特征图）的输出。在卷积操作时，卷积核被用于在图像（或特征图）上滑动以计算得到新的特征。可以看到，当卷积核滑动到图像（或特征图）的右上角时，输出特征对应位置（即第三个矩阵右上角位置）的像素值为卷积核窗口中所有像素的平均值。在实际计算卷积时，会对卷积核进行翻转。

为了更加灵活地提取特征，通常会引入卷积核的滑动步长（Stride）和填充（Padding）。卷积核在图像或特征图上的滑动是从左上方开始，按照从左至右、自上而下的准则进行，每次滑动的行数（或列数）称为滑动步长。填充则是指在图像或特征图的四周填充元素（通常是 0）。

如图 2-2 所示，下方的蓝色矩阵为输入的图像（或特征图），上方的绿色矩阵为卷积操作提取的特征。第一行为步长为 1 且不进行填充的卷积操作；第二行为步长为 2 且不进行填充的卷积操作，此时卷积核在图像（或特征图）上滑动时每次移动两个像素；第三行为步长为 1 而填充为 2 卷积操作，蓝色图像（或特征图）周围的虚线矩形即为填充的元素。

从图中也可以看到，卷积提取的特征图大小不仅与输入图像（或特征图）和卷积核的大小有关，而且与卷积核的滑动步长核填充元素的数量也有关系。这一关系即为式 2-2 和式 2-3。

$$W_{out} = \left\lfloor \frac{W_{in} + 2 \times p - w}{s} \right\rfloor + 1 \qquad (2-2)$$

$$H_{out} = \left\lfloor \frac{H_{in} + 2 \times p - h}{s} \right\rfloor + 1 \qquad (2-3)$$

式中，W_{out}，H_{out} —— 输出特征图的宽度和高度；

W_{in}，H_{in} —— 输入图像（或特征图）的宽度和高度；

w，h —— 卷积核的宽度和高度；

p，s —— 填充大小和步长。

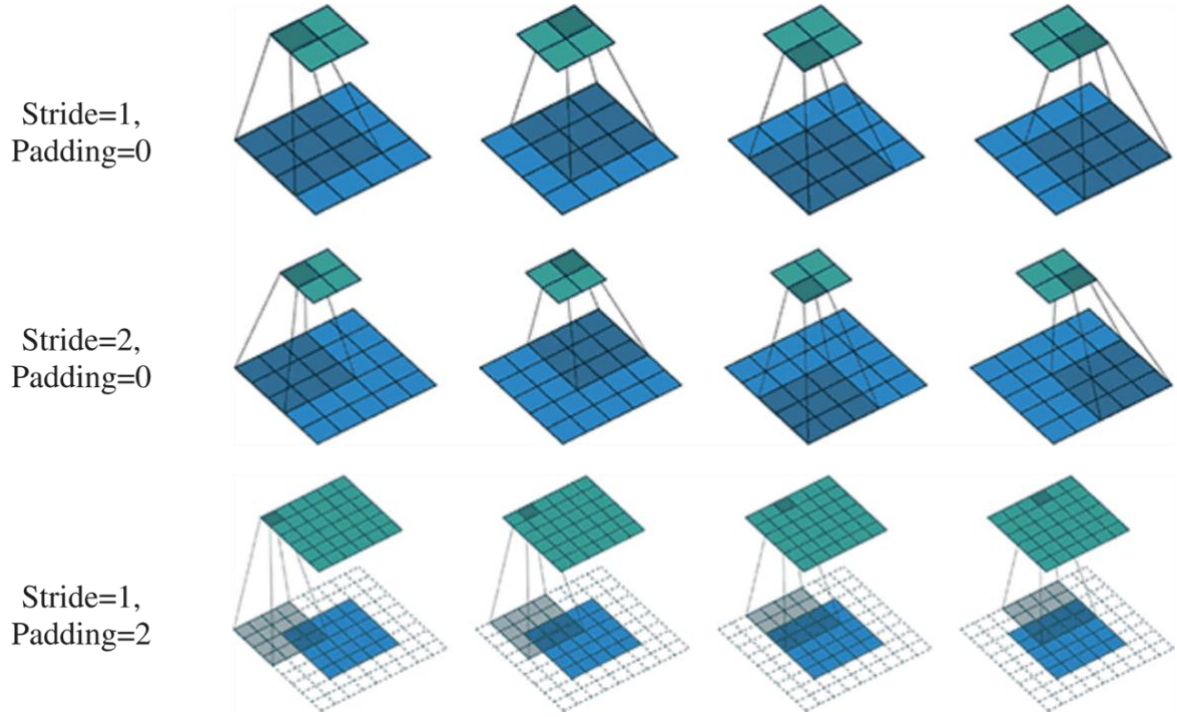

图 2-2 不同滑动步长和填充的卷积示例图

对于一幅图像来讲，除了目前已经提到的宽度和高度两个维度而言，还有颜色通道（例如，彩色图像有三个通道，深度图像只有深度一个通道），即给定一个图像（或特征图） $X \in R^{W \times H \times C}$ 和一个卷积核 $w \in R^{w \times h \times c}$，其中 C 和 c 分别为图像和卷积核的通道数，一般有 $C = c$。则卷积可以定义为对应通道的图像（或特征图）与卷积核进行卷积操作后相加，如图 2-3 所示。

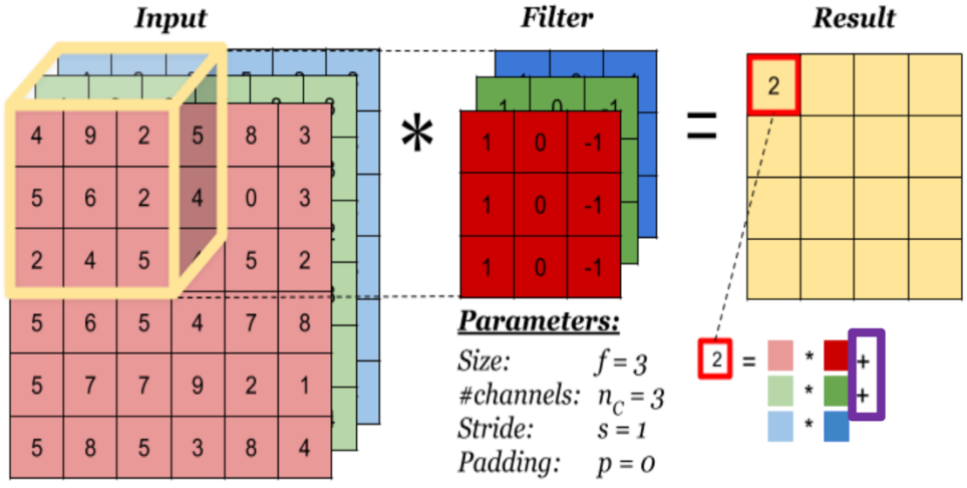

图 2-3 三通道彩色图像卷积示例图

图 2-3 中，左侧为输入的 RGB 三通道图像，中间为与图像通道数相同的卷积核，在对应颜色通道进行卷积操作后，将三个通道的结果相加（如图中右下角所示），即为输出特征对应位置的值。

2.1.2 激活函数

卷积操作在提取特征的过程中，能有效去除图像（或特征图）中的噪声。通过卷积层的堆叠，特征图的大小逐渐下降，感受野逐渐增大，特征包含的信息逐渐从低级的结构信息转变为高级的语义信息。但由于卷积操作的线性特征，即使是堆叠的卷积操作仍然面临着表示能力不足的问题。因而引入了激活函数这一非线性操作以增强模型的特征表示能力。此外，由于反向梯度传播在神经网络训练中取得的成功，决定了激活函数必须具备可导的特质。常用的激活函数有 Sigmoid， Tanh， ReLU， Leaky ReLU，其函数图像如图 2-4 所示。

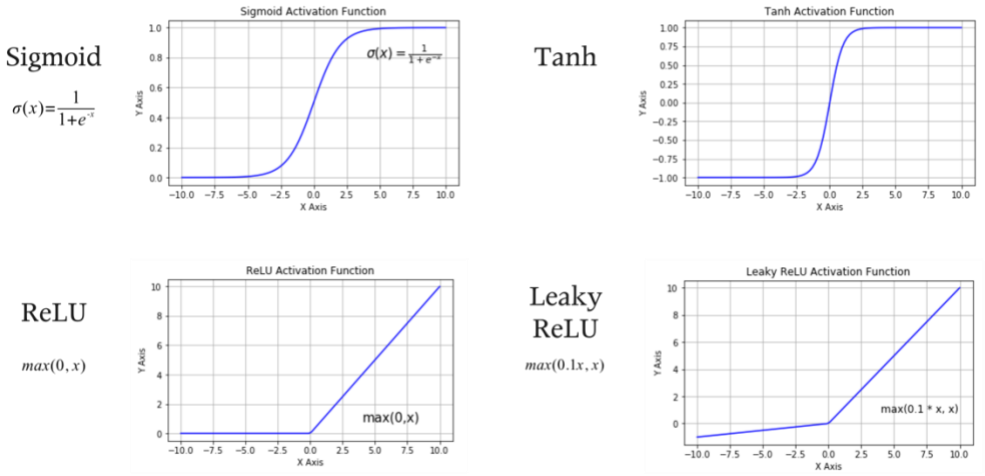

图 2-4 常用激活函数及其函数图像

2.1.3 池化

卷积层由于其稀疏连接的特性可以显著减少网络中连接的数量，但神经元的数量却并没有发生变化。可以预见的是，仅包含卷积层和激活函数的神经网络会导致分类器的输入维度很高。池化层（一种降采样）正是为解决这样的问题而被引入了卷积神经网络，以实现降低特征数量和减少参数量的效果。此外，卷积操作可以精确地找到像素变化的位置，但在实际的图像数据中，研究对象并不总是出现在固定位置，可能会导致相同的特征经过卷积后的输出不同，给下游任务带来不便。以人的视觉为例，我们可以从一张包含狗的降采样图像中识别出图像内容，这说明压缩后的图像仍保留了狗最关键的特征，而图像降质时丢掉的是一些无伤大雅的信息，最能描述狗的尺度不变特征则仍然保留在图像中。因此，池化层还有一个作用是特征选择，减少由于像素位置变化造成的特征变化。一般所使用的池化层为最大池化（Max-pooling）和平均池化(Average-pooling)两类。最大池化如图 2-5 所示。

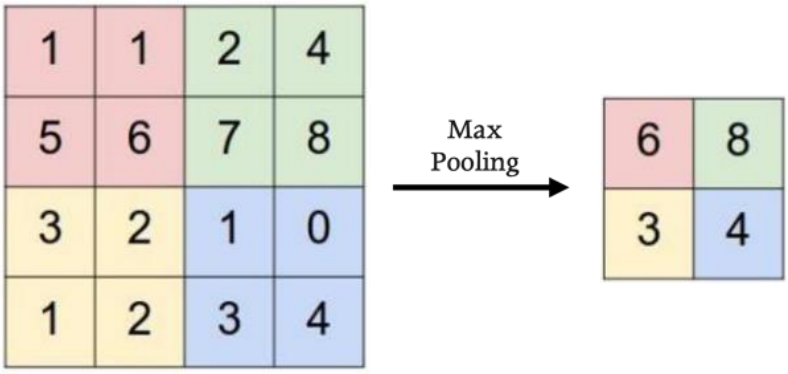

图 2-5 最大池化示例图

可以看到，最大池化是将输入的图像划分为若干个（图中为 4 个）矩形区域，然后分别取各个区域的最大值得到的。而平均池化则是取各个区域的平均值。

2.1.4 正则化

卷积神经网络在训练时通过误差的反向传播和梯度下降来不断更新网络的参数。而随着前面层在训练过程中对参数的更新，导致了传输到后面层作为输入数据的分布发生了变化。也就是说，在训练过程中，除了输入层之外，网络每一层输入数据的分布一直在发生变化，这一问题被称为内部协变量位移（Internal Covariate Shift）。而批正则化（或批量归一化）可以用于解决这一问题。批正则化可以对网络中任意一层进行归一化处理，保证每一层的数据都处于同一分布。这一算法可以定义为式 2-4。

$$\hat{x}^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\sqrt{Var[x^{(k)}]}} \qquad (2-4)$$

式中，$x^{(k)}$ —— 输入数据第 k 个维度的值；

$E\left[x^{(k)}\right]$ —— 第 k 个维度的期望；

$\sqrt{Var[x^{(k)}]}$ —— 第 k 个维度的方差。

2.1.5 全连接

综上，一般卷积神经网络的结构如图 2-6 所示。卷积神经网络中一般由 N 个连续的卷积块堆叠而成，最后有 F 个连续的全连接层（N 的取值可以很大，而 F 一般为 0～2）。其中，每个卷积块由 C 个卷积层（包括激活函数）和 P 个池化层（M 的取值通常为 2～5，而 b 为 0 或 1）串联组成。

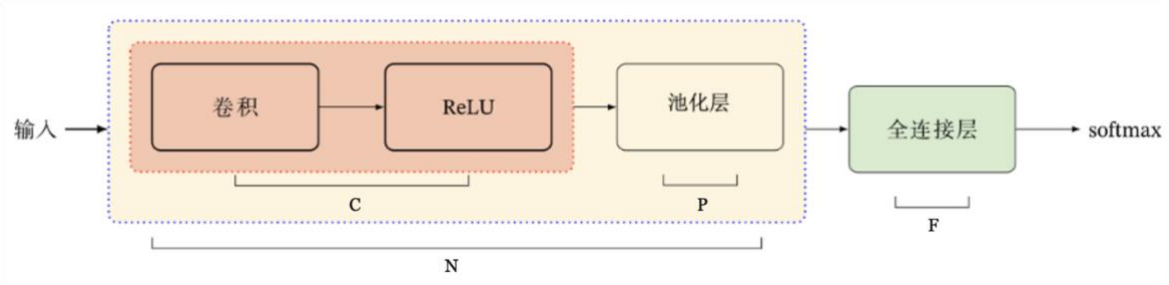

图 2-6 典型的卷积神经网络结构图

全连接层在卷积神经网络中一般用于分类。在通过连续的卷积层和激活函数捕捉到了原始图像的特征表示后，全连接层将提取的特征聚合并转换到最终的样本空间，如图 2-7 所示。

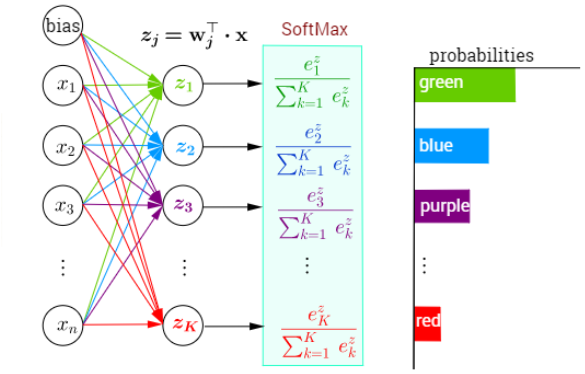

图 2-7 全连接层示例图

顾名思义，与卷积的稀疏连接不通，从图 2-7 可以看到，全连接层中的每一个“神经元”都与上一层中的所有“神经元”相连。全连接层将网络提取的二维图像特征映射为样本标记类别得分（取值范围为 $[-\infty, +\infty]$）的一维向量，而通常网络的全连接层后会紧跟一个 Softmax 操作，将类别得分转换为概率以便于后续操作。 Softmax 操作的定义为式 2-5。

$$\hat{y}_i = softmax(z_i) = \frac{e^{z_i}}{\sum_K e^{z_i}} \qquad (2-5)$$

式中，z_i，$\hat{y}_i$ —— 第 i 个类别的得分和对应的概率；

K —— 分类的类别数或全连接层的输出维度。

2.1.6 其他卷积

前文介绍的卷积实现了低层特征到高层特征的转换，同样，也可以实现特征维度间的转换。例如，通过三个卷积核可以实现将通道数为 5 的特征转变为通道数为 3 的特征。而在一些明确的任务（如图像的超分辨率重建）中，存在着将特征从低层到高层进行转换的需要，也就是对特征图或图像进行“上采样”。转置卷积（Transposed Convolution）

也被称为反卷积（Deconvolution），通过填充扩大特征，实现了由低分辨率特征图到高分辨率特征图的操作。如图 2-8（右）所示。

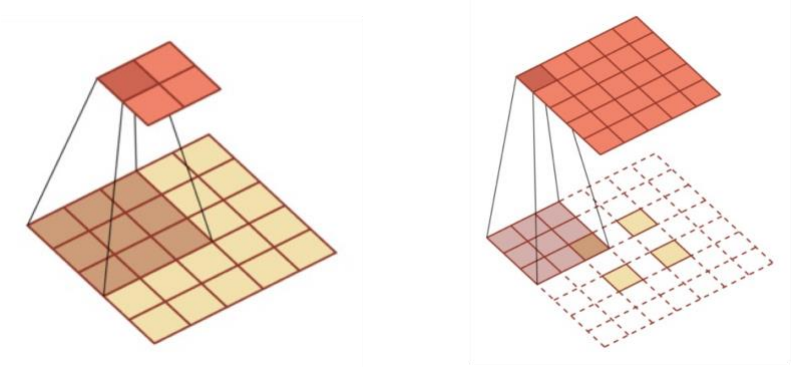

图 2-8 转置卷积（反卷积）示例图

图中下方的黄色矩阵为输入图像（或特征图），上方的橙红色矩阵为“卷积”操作的输出。可以看到与左侧卷积操作不同的是，反卷积的输出比输入图像（或特征）在空间维度上大，从而可以承担图像超分辨率重建任务中上采样的工作。

2.2 编码器-解码器框架

编码器-解码器（Encoder-Decoder）框架最先被应用于自然语言处理领域的有关任务，如机器翻译等。编码器是以训练数据作为输入，特征张量作为输出的网络。这些特征张量就是输入的信息和特征嵌入到其他表示空间的内容。解码器通常是一个与编码器结构相似（或与特定任务有关的结构）但方向相反的网络结构，它从编码器获取特征张量，然后输出与实际输入或预期输出同一表示空间的结果。近年来，编码器-解码器架构作为一类深度学习框架，也被广泛应用于计算机视觉领域，并在图像分割、光流估计、深度估计等任务取得了成功。也就是说，编码器的输入可以是文字，图像或视频数据，而具体的模型可以采用卷积神经网络，循环神经网络等。以卷积神经网络作为编码器和解码器的模型通常如图 2-9 所示，左侧为网络的编码器，以彩色图像作为输入并生成图像的特征，右侧为网络的解码器，以编码器提取的特征作为输入，然后将特征映射到预期输出的表示空间，这里为图像的语义标签。

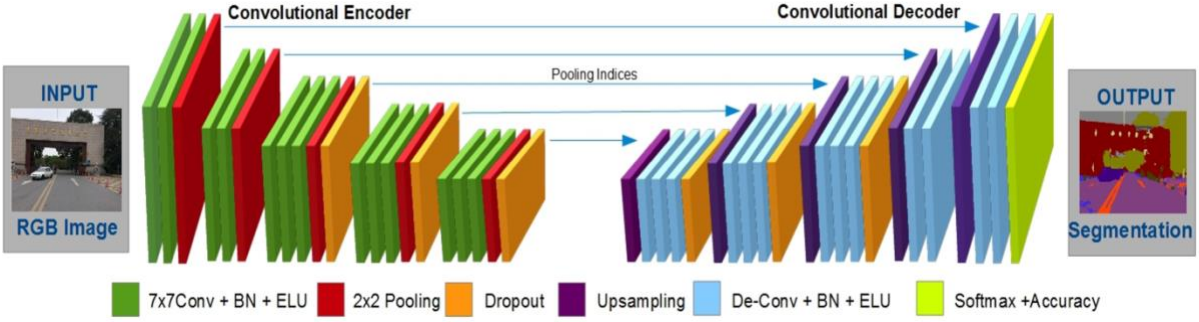

图 2-9 编码器-解码器框架示例图

2.3 注意力机制

在观看一张图片时，人们并不能立即接收图片的全部内容，而是往往将注意力集中在了一些显著位置。图 2-10 所示的是对人浏览报纸时注意力的模拟，可以看到注意力集中在图像中的显眼目标（如标题）或直观且包含大量信息（如图片）的位置。

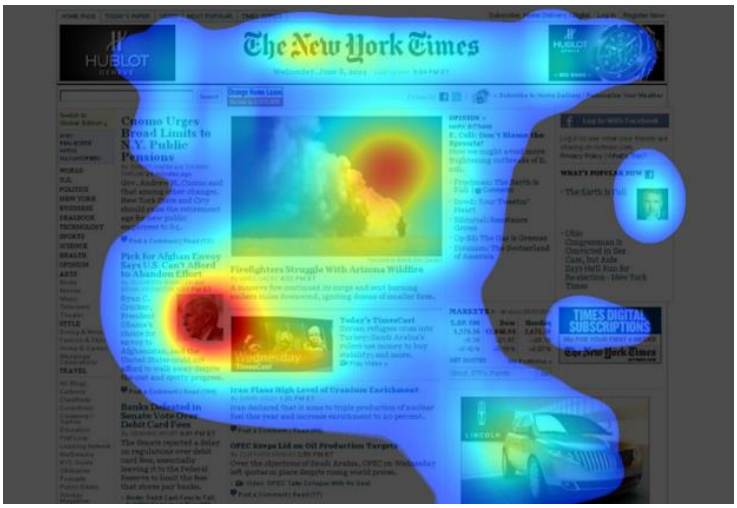

图 2-10 注意力示例图

同理，注意力机制的作用便是帮助模型对输入数据赋予不同的权重，使得模型关注输入数据中更加重要和关键的信息。此外，注意力机制还可以使得模型在做出准确判断的同时而不带来巨大的计算和存储开销。

基于项的注意力和基于位置的注意力根据其不同的输入形式而有所不同。基于项的注意力用于在输入中已经给出或可以提取得到项（如向量，矩阵等）的任务。而基于位置的注意力则用于从输入难以获取项的任务，因而一般以特征图作为输入，或在输入前对该特征图进行一定的变换使得网络在后续阶段可以更好地利用它。

根据学习机制的不同，注意力可以分为软注意力和强注意力。软注意力通常关注区域或通道，且会对所有输入进行线性组合。因此，软注意力一般由端到端的结构（或网络）生成，即可以通过梯度下降的方法学习得到。而硬注意力则是根据输入数据的分布随机进行选择，也就是说，强注意力是不可微的。强注意力的训练过程一般是通过强化学习（Reinforcement Learning） 的收益函数（Reward）来进行激励，从而让模型更加关注数据的局部细节。

而按照注意力域的不同，注意力又可以分为通道注意力（Channel Attention），空间注意力（Spatial Attention）和混合注意力（Mixed Attention）。下面将分别对使用通道注意力和空间注意力思想的经典网络结构进行介绍。

2.3.1 通道注意力

卷积神经网络通过卷积等操作提取包含空间和通道信息的特征，现有很多针对网络的研究是从增强特征的空间表示能力的角度来提升网络的性能。而通道注意力则是对特征通道间的关系进行建模，让网络可以自动地学习不同通道的重要性。压缩-激励网络(Squeeze-and-Excitation Network, SENet)[27]是一个利用通道注意力思想的经典网络。如图 2-11 所示，其操作是对特征图进行空间维度的压缩，从而将具有通道级特征响应的全局分布嵌入一个一维向量中，并通过激励操作获得每个通道权重的集合，然后分别与对应通道相乘，从而得到施加注意力的结果。

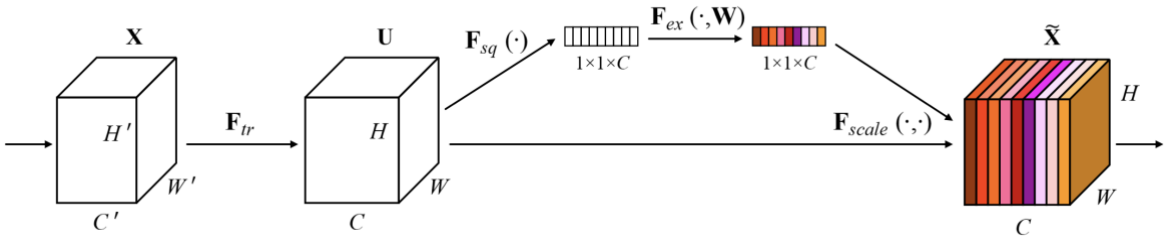

图 2-11 压缩-激励模块结构图

2.3.2 空间注意力

对于卷积神经网络所提取的特征图 $F \in R^{W\times H\times C}$，$W$，$H$，$C$ 分别代表特征图的宽度、高度和通道数量。$W \times H$ 可以看作一个平面，其中的每个值代表的当前位置的信息。尽管经过卷积等操作，这个位置已经不再是原始图像像素的位置，但是依然是一种位置信息。与通道注意力类似，如果可以学习到一个 $W \times H$ 的权重矩阵，将矩阵的每一个元素与特征图所有通道对应位置的元素相乘，就相当于对空间位置进行了加权，即对特征图施加了空间注意力。换言之，空间注意力是对特征图的每个位置进行加权调整，促使模型关注到更加值得关注的区域。

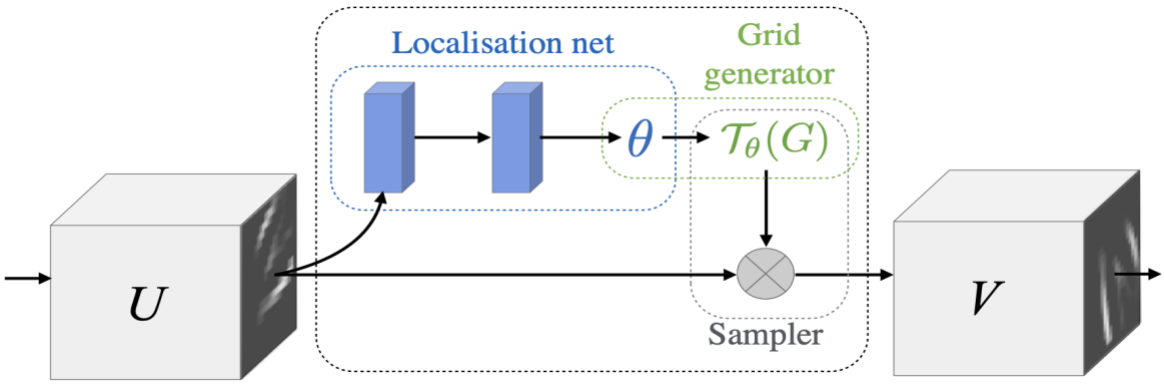

图 2-12 空间转换模块结构图

空间转换网络（Spatial Transformer Networks, STN）[28]是一个利用空间注意力思想的经典网络，其动机是通过设计的模块让网络学习一种变换，这种变换可以将进行了仿射变换的目标进行矫正。通俗而言，就是将图像或特征的空间信息做对应的变换，从而将关键信息提取出来，即找出图片信息中需要被关注的区域。如图 2-12，这种变换的学习主要是通过局部网络（Localisation Network）、参数化网格采样（Parameterised Sampling Grid）和差分图像采样（Differentiable Image Sampling）实现的，分别用于变换参数的预测、变换前后图片坐标的映射和像素的采集。

2.4 多尺度机制

在计算机视觉任务中，尺度微小的物体与超大的物体对模型的性能有着很大的影响。目标特征在图像逐层经过卷积神经网络的过程中被提取出来，卷积层的感受野也随着网络的深入不断变大。但感受野过大，提取的特征会包含冗余的无效信息，而感受野太小，就只能捕获到图像的局部特征。为了实现精确的目标检测、识别或是像素级回归，就需要多尺度技术来增强网络的特征表示能力。多尺度机制就是对图像进行不同粒度的采样，模型在不同的尺度下可以捕捉到不同的特征。根据网络设计的不同可以分为多尺度输入、多尺度特征融合和多尺度预测融合。

多尺度输入就是将不同尺度的图像作为输入，也就是图像金字塔。多尺度特征融合通常有两种结构，一种是多分支并行的结构（如 Inception 模块），通过多分支在同一层提取不同粒度大小的特征，融合后传递到网络的下一层。并行的结构可以更加灵活地平衡计算量，并提高模型的特征表示能力。另一种是跨层连接的串行结构（如全卷积神经网络），这种结构的实现方式是通过跳连接将不同层的特征进行融合。串行的结构可以帮助边界敏感任务（如语义分割）更好地捕捉边界细节。图 2-13（左）为并行分支的多尺度特征融合结构，图 2-13（右）为串行跳层连接的多尺度特征融合结构。

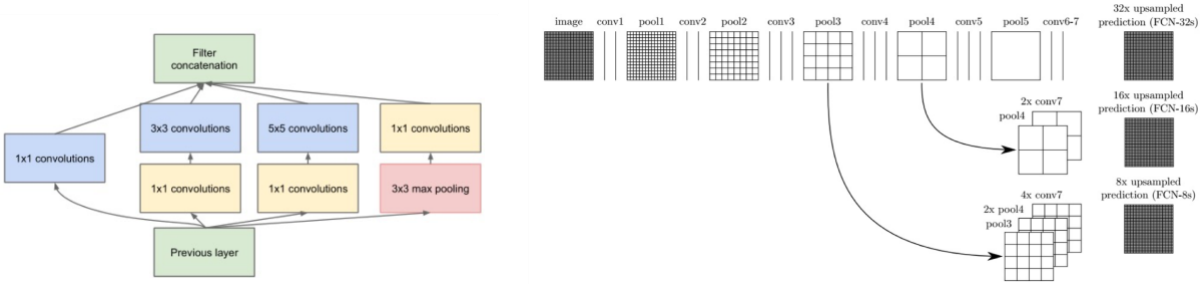

图 2-13 多尺度特征融合结构示例图

多尺度预测融合是根据不同尺度的特征进行预测，然后将不同尺度的结果融合。但多尺度输入结构独立地在不同尺度的图片上进行计算，造成了巨大的计算开销，而多尺度特征融合与多尺度预测融合结构由于缺乏对特征金字塔的充分利用，其准确度仍有待

提升。特征金字塔网络则是结合了多尺度特征融合与多尺度预测融合的特点，在每个分辨率的特征图与前一层上采样后的特征图逐像素相加，从而使得每一层用于预测的特征图融合了不同分辨率和不同语义信息的特征。与其他的多尺度方法相比，特征金字塔的每一层的特征都具有合适的分辨率以及丰富的语义。特征金字塔网络的结构如图 2-14 所示。

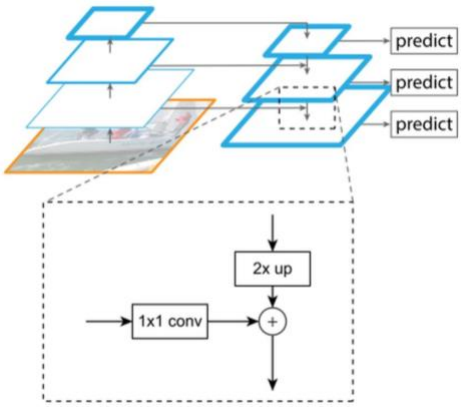

图 2-14 特征金字塔网络结构图

2.5 像素重组

像素重组（PixelShuffle）以低分辨的特征图左为输入，通过卷积和特征通道间的重组得到高分辨率的特征图，从而可以有效地解决图像的超分辨率重建问题。如图 2-15 所示，在隐藏层中进行的是对图像的特征提取，紧随其后的一层生成 $w \times h \times r^2$ 的特征图，其中，w 、h、r^2 分别为特征图的宽度、高度和通道数量，r 为图像超分辨率重建中的上采样因子。像素重组作用于这一特征图并重新组合为宽度和高度分别为 $w \times \mathrm{r}$，$r \times h$ 的上采样特征图。具体而言，像素重组就是把低分辨特征图的一个像素划分为 r 个“亚像素”，然后将特征图 r 个通道对应位置的值按照预先制定的规则来填充划分出的“亚像素”。在每个低分辨特征图像素划分出的“亚像素”被填满后，就完成了像素重组。在填充过程中，像素重组层可以更新 r^2 个通道的权重以优化生成的高分辨率特征图。

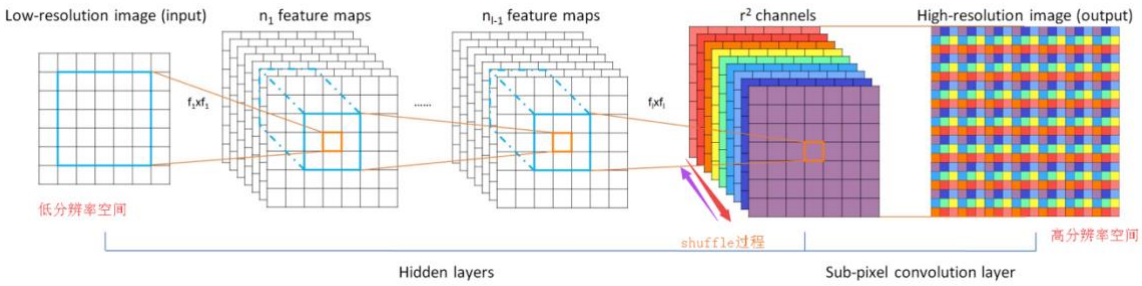

图 2-15 像素重组结构图

2.6 本章小结

本章介绍了本文所提出的单目深度估计和深度图像超分辨率重建联合学习网络所使用的技术和方法。本章所介绍的内容主要分为计算机视觉任务的通用方法和与图像超分辨率重建任务密切相关的方法，其中计算机视觉任务通用方法即为卷积神经网络。而在后续介绍与图像超分辨率重建任务相关的技术时，如多尺度机制，像素重组等，则是围绕着这些技术与卷积神经网络的相关性来引入和展开的。也就是说，卷积神经网络是计算机视觉任务的技术基石，而其他相关理论和方法则是推动细分任务发展的重要工具。综上，本文提出的单目深度估计和深度图像超分辨率重建联合学习网络整体上采用编码器-解码器框架，以卷积神经网络作为网络设计的原型，辅助以注意力机制、多尺度机制和像素重组等方法，从而构成了完整的网络设计。

3 算法实现

第一章的分析指出，现有研究已经证明了单目深度估计在改善深度图像超分辨率重建性能方面的有效性，但不足的是并没有明确地探究两个任务的相关性。与之前的工作不同，本文提出了单目深度估计和深度图像超分辨率重建的联合学习网络（BridgeNet），以实现更好的深度图像超分辨率重建性能。

本文提出的联合学习网络是基于第二章所陈述的理论和技术构建的，本章将按照自顶向下的逻辑，由网络的整体到局部分层进行介绍。首先将介绍本文提出的单目深度估计和深度图像超分辨率重建的联合学习网络（BridgeNet）的网络架构及训练策略，然后将介绍用于子网络间交互的桥接器的设计动机和模块功能。

3.1 网络架构

图 3-1 描述了本文提出的单目深度估计和深度图像超分辨率重建联合学习网络（BridgeNet）的总体架构，该网络由两个子网络（即深度图像超分辨率重建子网络和单目深度估计子网络）和两个桥接器（即高频注意力桥和内容引导桥）组成。将深度图像超分辨率重建子网络（DSRNet）和单目深度估计子网络（MDENet）集成到一个统一的框架中，以实现深度图像超分辨率重建和单目深度估计的联合学习，并分别将高频注意力桥（HABdg）和内容引导桥（CGBdg）应用于网络的编码器和解码器，以将这两个任务桥接在一起。

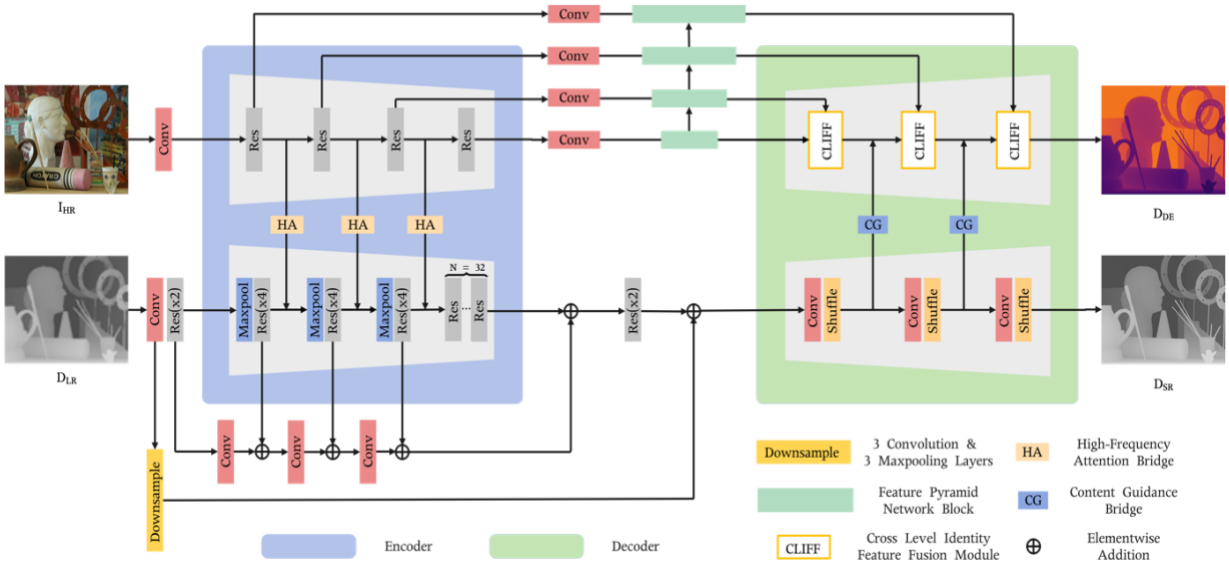

图 3-1 单目深度估计和深度图像超分辨率重建联合学习网络（BridgeNet）架构

给定一组高分辨率的 RGB-D 图像对 $\{I_{HR}^{(n)}, D_{HR}^{(n)}\}_{n=1}^{N}$ 和相应的低分辨率深度图像 $\{D_{LR}^{(n)}\}_{n=1}^{N}$ 作为训练数据，其中 N 是训练图像的数量。此外，低分辨率深度图像在输入网络前被插值到高分辨率深度图像的大小。本文提出的网络以低分辨率深度图像（D_{LR}）和相应的高分辨率彩色图像（I_{HR}）作为输入，同时对深度图像超分辨率重建子网络和单目深度估计子网络进行训练。超分辨的深度图像（D_{SR}）是本文网络的主要输出，此外，估计的深度图像（D_{DE}）也作为辅助输出。

3.1.1 单目深度估计子网络

编码器-解码器（Encoder-Decoder）结构在单目深度估计任务上取得了巨大的成功。本文遵循现有工作[29]中使用的编码器-解码器网络结构作为本文联合学习网络的单目深度估计子网络，该子网络由三个部分组成，即特征提取器，特征金字塔网络和深度预测模块。

特征提取器从输入的彩色图像中提取出多种分辨率的多级特征图的集合。然后，通过横向连接将生成的特征图馈送到特征金字塔网络中。特征金字塔网络将提取的语义信息从高层特征传播到低层特征，进而生成优化后的多级特征，即特征金字塔。深度预测模块基于特征金字塔完成最终的深度估计。本文采用经过预训练的 ResNet-50 网络的前 5 层作为单目深度估计子网络的特征提取器。

为了充分利用特征金字塔，现有的一些方法采用直接融合的策略。在这种策略的指导下，首先将特征金字塔中的所有特征图上采样到相同的分辨率，然后级联在一起用于估计深度图像。尽管具有丰富语义信息的高层特征所估计的深度图像鲁棒性更强，但从非常低的分辨率直接对它们进行上采样，会由于上采样特征图的模糊而导致所估计的深度图像也很模糊。

另一种策略是对特征金字塔中不同分辨率的特征渐进式地融合。这种方法首先对高层特征进行逐步上采样，然后与相同分辨率的低层特征进行融合。虽然这样的融合方式可以在一定程度上抑制在深度估计时的模糊问题，但其输出特征主要受低级特征所决定，而这样的特征对于较难估计的场景并不具有鲁棒性。

在本文的单目深度估计子网络特征解码期间，跨层恒等特征融合（Cross-level Identity Feature Fusion, CLIFF）模块用于逐步融合优化的多层特征并完成深度估计，其结构如图 3-2 所示。跨层恒等特征融合模块以高层特征图和低层特征图作为输入。首先使用双线性插值（Bilinear Interpolation）将高层特征图进行上采样，使得输入的两个特征图具有相同的分辨率。由于高层特征具有丰富的语义信息且噪声更少，因此借用注意力机制的思想，通过将低层特征与高层特征相乘来对低层特征进行强化。如此一来，低层特征中的准确响应得到进一步的增强，而噪声响应则会受到抑制。为了得到高层特征、

原始的和强化后的低层特征的最佳组合，跨层恒等特征融合模块通过两个卷积层来对三个特征进行进一步选择。

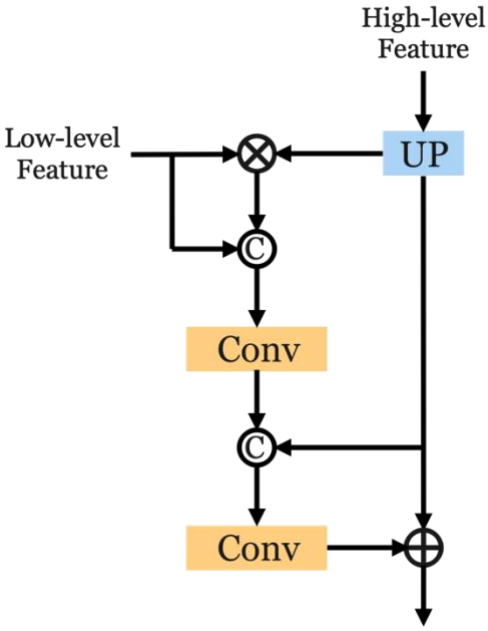

图 3-2 跨层恒等特征融合模块结构图

具体而言，原始的低层特征和强化后的低层特征在级联后将作为第一个卷积层的输入，而第一个卷积层用于学习从级联的特征中提取更有用的信息。其输出将与高级特征级联起来，用作第二个卷积层的输入，从而实现对低层特征和高层特征的选择与聚合。此外，出于保留高级语义特征的考虑，在上采样的高层特征与第二个卷积层的输出特征之间添加了恒等映射（Identity Mapping）。通过跨层恒等特征融合模块得到的特征，兼顾了高层特征的语义信息和低层特征的结构信息，是对单目深度估计最有利的特征组合。

跨层恒等特征融合模块中所进行的操作可以描述为式 3-1 至式 3-4。

$$F^a = F^l \odot F^h \tag{3-1}$$

$$F_1^c = W_1([F^l, F_a]) \tag{3-2}$$

$$F_2^c = W_2([F_1^c, F^h]) \tag{3-3}$$

$$F^o = F_2^c + F^h \tag{3-4}$$

式中，F^l，F^a，F^h —— 原始的和强化后的低层特征及高层特征；

F_1^c，F_2^c，F^o —— 两个卷积层的输出及最终输出；

W_1，W_2 —— 两个卷积层的权重；

$\odot$ —— 像素级相乘；

$[\cdot,\cdot]$ —— 级联操作。

3.1.2 深度图像超分辨率重建子网络

遵循人脸超分辨率网络的结构[30]，本文同样将编码器-解码器网络作为深度图像超分辨率重建的基线。其中，编码器由一个浅层特征提取器（包含一个卷积层和一个残差块）和三个连续的特征编码模块以及一些堆叠的残差块构成。每个特征编码模块由一个最大池化层和四个串联的残差块组成，残差块的结构如图 3-3 所示。最大池化层可以确保深度图像超分辨率重建子网络编码器每一层的特征与单目深度估计子网络编码器对应层特征的分辨率相同，以便后续特征指导的顺利进行。

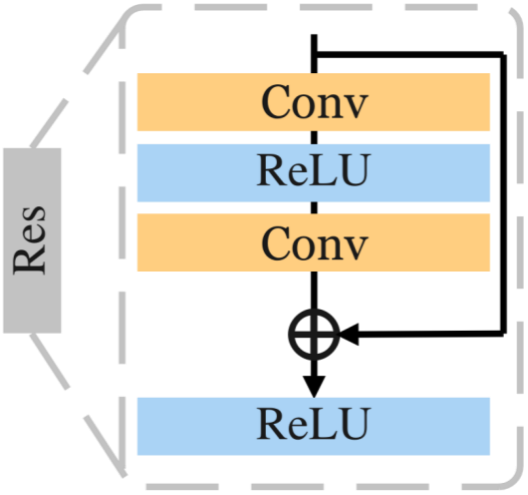

图 3-3 残差模块结构图

在深度图像超分辨率重建子网络中，所有卷积层都使用大小相同的 3×3 卷积核，且每个卷积操作后面紧随一个 ReLU 激活函数。除最后的卷积层通道数设置为 3 外，子网络中其余的所有卷积操作的通道数均设置为 256。深度图像超分辨率重建子网络所采用的残差模块与原始残差网络的结构基本一致，不同的是移除了残差块中的批量归一化层。

这样做的动机是由于批量归一化层（正则化层）对特征进行了归一化处理，从而导致网络丢弃了范围灵活性，所以最好是移除这些批量归一化层。现有研究[31]的实验也表明用于图像超分辨率重建的残差网络在去除所有的批量归一化层后表现最佳。不仅如此，批量归一化层会减慢网络收敛的速度，同时降低其整体性能，尤其是在图像的超分辨率重建任务中。因此在深度图像超分辨率重建子网络中，所有残差模块中的批量归一化层都被移除。

更深层的网络已经在包括图像超分辨率重建在内的许多计算机视觉任务中显示出更好的性能。增加网络的深度也是现有许多工作中使用的一种策略。经过三个特征编码模块提取出的特征将传递到由 N 个残差块组成的更深层的特征提取模块中。较深的网络可以帮助超分辨的深度图像恢复出更分明的边缘和形状。

为了恢复深度图像中的精细结构和微小物体，本文引入了多尺度策略，通过进一步融合深度图像超分辨率重建子网络编码器中间层的特征来优化高层特征。考虑到相邻层特征的相似性，并非编码器的所有特征都需要参与多尺度特征融合以形成新的特征表示，而是仅仅采用了每个特征编码模块提取的特征进行融合。因此每个最大池化层之前的特征编码模块的输出都需要通过跳连接，与深层特征模块输出的特征融合以形成具有更丰富信息的特征。此外，由于特征图每经过一个最大池化层都进行了 2 倍的下采样，为了匹配不同层特征图的大小，将步长为 2 的 3×3 卷积层应用于由跳连接提供的中间层特征，以实现下采样并进行特征融合。

对于低分辨率的深度图像 D_{LR}，上述操作可以被描述为式 3-5 至式 3-7。

$$F_0 = f_0(D_{LR}) \tag{3-5}$$

$$F_i = f_i(F_{i-1}), i \in \{1,2,3\} \tag{3-6}$$

式中，F_0 —— 通过一个卷积层和一个残差块提取的浅层特征；

$F_i, i \in \{1,2,3\}$ —— 特征编码模块的输出；

f_0 —— 浅层特征提取模块；

$f_i, i \in \{1,2,3\}$ —— 特征编码模块。

$$F_{fusion} = g_3(g_2(g_1(F_1) + F_0) + F_2) + F_3 \tag{3-7}$$

式中，F_{fusion} —— 多尺度特征融合的输出；

$g_i, i \in \{1,2,3\}$ —— 实现对中间层特征下采样的卷积层。

在特征解码期间，本文使用了三个顺序连接的相同上采样模块，包括一个卷积层和一个像素重组层（PixelShuffle）。特征图每经过一个上采样模块后，分辨率都变为之前的两倍。最后应用一个 1×1 的卷积层来重建高分辨率的深度图像。

受彩色图像超分辨率重建算法稠密残差网络（Residual Dense Network, RDN）[32]的启发，编码器初始阶段提取的浅层特征和输入解码器的特征在进行第一次上采样前通过一个大的跳连接进行融合。设计这个跳连接的目的是为了直接提供深度图像超分辨率重建需要的低频信息，从而迫使网络更加专注于学习重建需要的高频信息而不是已经提供的低频信息。

根据网络的设计可以看到，网络提取的浅层特征图大小与上采样后的低分辨率特征图大小相同。而网络中有 3 个最大池化层，每个最大池化层对特征图进行因子为 $\times 2$ 的下采样，即特征图在经过编码器中三个最大池化后，分辨率下降为输入图像的 1/8。为了保证长跳连接两侧的特征图拥有相同的分辨率，浅层特征被馈送到下采样（Downsample）块中以生成低频特征，然后通过长跳连接与编码器提取的最终特征相融合，下采样块的结构如图 3-4 所示。

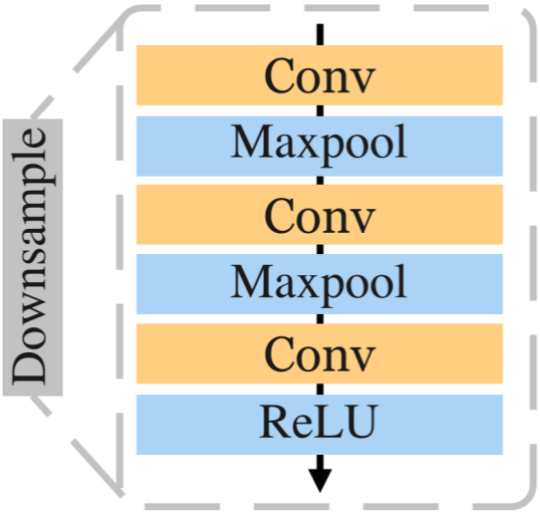

图 3-4 下采样模块结构图

3.1.3 联合学习策略

深度图像超分辨率重建和单目深度估计具有天然的相关性，可以在同一数据集的监督下对他们进行训练。因此，这两个任务的联合学习首先体现在损失函数的联合优化上。

与其他多任务学习的损失函数为所有分支损失函数的加权和不同，本文分别为深度图像超分辨率重建和单目深度估计的损失函数分配了不同的优化器。这是因为深度图像超分辨率重建和单目深度估计的学习难度大不相同，导致两个任务的收敛速度不同，从而很难找到合适的权重设置来确保两个任务都达到最佳性能。因此，在损失函数的设计方面，本文提出分别对深度图像超分辨率重建和单目深度估计相关部分进行优化的策略。其损失函数定义为式 3-8 和式 3-9。

$$\mathcal{L}_{DSR} = \|D_{SR} - D_{HR}\|_1 \quad (3-8)$$

$$\mathcal{L}_{MDE} = \|D_{DE} - D_{HR}\|_1 \quad (3-9)$$

式中，$\mathcal{L}_{DSR}$ —— 深度图像超分辨率重建任务的逐像素 L_1 损失;

$\mathcal{L}_{MDE}$ —— 单目深度估计任务的逐像素 L_1 损失。

除了在损失函数层级对两个任务进行约束外，本文还精心设计了两个桥接器，分别将两个任务在编码器和解码器阶段相关联，以实现互利共赢。一个是编码器中的高频注意力桥（HABdg），另一个是解码器中的内容引导桥（CGBdg）。其中，高频注意力桥利用单目深度估计子网络出色的彩色特征学习能力为深度图像超分辨率重建提供指导。考虑到两个任务难度的差异，深度图像超分辨率重建子网络在解码器阶段经由内容引导桥可以为深度估计提供有效的内容指导，从而提高深度估计的性能。在以下各小节中，本文将详细介绍这两个桥接器的设计原理和实现细节。

下面给出深度图像超分辨率重建和单目深度估计联合学习网络训练策略的伪代码。

Algorithm 1: Joint Learning Strategy

Input: Training data D_{LR}, I_{HR}, D_{HR}

Output: D_{SR}, D_{DE}

Randomly initialize DSRNet and MDENet

for *epoch=1; epoch* $\leq$ *400;* **do**

——— Step 1 ———

$F_{MDE} = Encoder_{MDE}(I^n_{HR})$

$F^{shallow}_{DSR} = Res^{(2)}(conv(D_{LR}))$; // shallow feature extraction

for i=1; $i \leq 3$ **do** // i refers to i^{th} layer of encoder

  if i=1 **then**

    $Fe^i_{DSR} = maxpool(Res^{(4)}(F^{shallow}_{DSR}))$

  else

    $Fe^i_{DSR} = maxpool(Res^{(4)}(F^{i-1}_{ha}))$

  $F^i_{blurred} = deconv(avgpool(F^i_{MDE})$

  $A^i_{hf} = PRelu(F^i_{MDE} - F^i_{blurred})$

  $F^i_{hg} = F^i_{MDE} + A^i_{hf} \cdot F^i_{MDE}$

  $F^i_{comp} = [Fe^i_{DSR}, F^i_{hg}]$

  $F^i_{ha} = SA(conv_{1\times1}(CA(F^i_{comp})))$

$F^{deeper}_{DSR} = Res^{(32)}(F^3_{ha})$

$F^{multi-scale}_{DSR} = conv(conv(conv(Fe^{shallow}_{DSR} + Fe^1_{DSR}) + F^2_{DSR}) + Fe^3_{DSR})$; // multi-scale features fusion

$F^{fusion}_{DSR} = Res^{(2)}(F^{deeper}_{DSR} + F^{multi-scale}_{DSR})$

$F^{low-freq}_{DSR} = Downsample(F^{shallow}_{DSR})$

for i=1; $i \leq 3$ **do** // j refers to j^{th} layer of decoder

  if i=1 **then**

    $Fd^j_{DSR} = pixelshuffle(conv(F^{fusion}_{DSR} + F^{low-freq}_{DSR}))$

  else

    $Fd^j_{DSR} = pixelshuffle(conv(F^{j-1}_{DSR}))$

$D_{SR} = conv_{1\times1}(Fd^3_{DSR})$

Update weights of parts related to DSR with $\mathcal{L}_{DSR} = ||D_{SR} - D_{HR}||_1$

——— Step 2 ———

$F_{MDE} = Encoder_{MDE}(I^n_{HR})$

for k=1;$k \leq 4$ **do**

  if $Fp^k_{MDE} = conv(Fe^{5-k}_{MDE})$ **then** k=1

  $Fe^{5-k}_{MDE} = interpolate(Fe^{5-k}_{MDE})$

  $Fp^k_{MDE} = conv(Fe^{5-k}_{MDE} + Fe^{5-(k+1)}_{MDE})$

for j=1;$k \leq 3$ **do**

  if j=1 **then**

    $Fd^j_{MDE} = CLIFF(Fp^1_{MDE}, Fp^2_{MDE})$

  else

    $M^j_{DSR} = conv_{1\times1}(Fd^j_{DSR})$

    $M^j_{MDE} = conv_{1\times1}(Fd^j_{MDE})$

    $W^j_{diff} = softmax(conv_{1\times1}(M^j_{DSR} - M^j_{MDE}))$

    $F^j_{ca} = Fd^j_{DSR} + W^j_{diff} * Fd^j_{DSR}$

    $F^j_{con} = [Fd^j_{MDE}, F^j_{ca}]$

    $F^j_{cg} = SA(conv_{1\times1}(CA(F^j_{con})))$

    $Fd^j_{MDE} = CLIFF(F^j_{cg}, Fp^{j+1}_{MDE})$

$D_{DE} = interpolate(conv_{1\times1}(Fd^3_{MDE}))$

Update weights of parts related to MDE with $\mathcal{L}_{MDE} = ||D_{DE} - D_{HR}||_1$

3.2 高频注意力桥

回顾现有的颜色指导的深度图像超分辨率重建的方法可以发现，彩色图像的指导主要包括对应特征的直接引导[10, 12, 16]或边缘细节的引导[33-34]。尽管彩色图像和深度图像具有很强的结构相似性，但是彩色图像丰富的纹理和边缘并不总是与深度图像一致，因此直接的特征引导或边缘引导可能会导致纹理复制和深度流失等问题。

在联合学习视角下，单目深度估计为解决上述问题提供了新的思路。单目深度估计任务是以彩色图像作为输入，然后实现将场景从光度表示映射到几何表示，从而生成深度图像。因此，由单目深度估计编码器提供的彩色图像的特征更接近于深度模态的特征表示。由单目深度估计编码器为深度图像超分辨率重建任务提供高频信息指导时，可以避免明显的伪影。这就是本文提出在编码器阶段使用单目深度估计代替现有方法的颜色分支来指导深度图像超分辨率重建的原因。

在明确了指导信息的传递方向之后，接下来需要思考的问题便是如何有效地实现。最简单直观的方法是通过级联或相加将单目深度估计子网络相应层的特征直接传递到深度图像超分辨率重建子网络中，但这显然不是明智的方法。在单目深度估计子网络的编码器中，随着网络的深入，特征图的分辨率逐渐降低，其中高层特征具有丰富的语义信息，而低层特征则具有更多的结构信息。由于低分辨率深度图像包含的高频信息较少，因此高分辨率的彩色图像可以为深度图像超分辨率重建提供更为重要的高频信息（例如边缘细节），而不是图像的语义信息。因此，本文设计了一个高频注意力桥，该桥接器用于从单目深度估计子网络学习高频信息以用于指导深度图像超分辨率重建子网络。高频注意力桥的结构如图 3-5 所示。

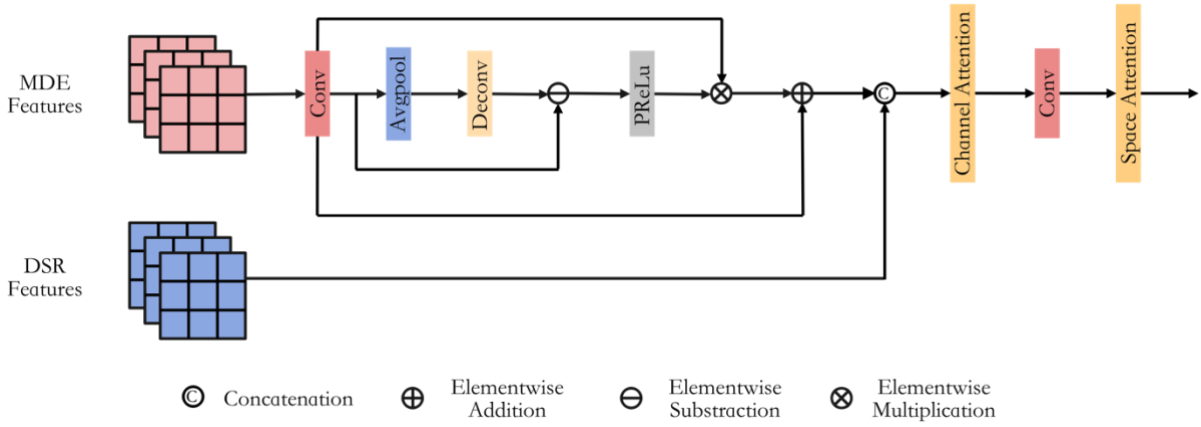

图 3-5 高频注意力桥结构图

具体来说，首先使用平均池化和反卷积运算对单目深度估计子网络的原特征进行模糊操作，这一操作可以被描述为式 3-10。

$$F^{i}_{blurred} = deconv\left(avgpool\left(F^{i}_{MDE}\right)\right) \tag{3-10}$$

式中，$F_{blurred}^{i}$ —— 获得的单目深度估计子网络第 i 层的模糊特征；

F_{MDE}^{i} —— 单目深度估计子网络第 i 层的特征；

$avgpool(\cdot)$ —— 平均池化操作；

$deconv(\cdot)$ —— 反卷积操作。

然后，通过将原始特征与模糊特征相减来获得特征中的高频信息，进而生成高频信息的注意力，即式 3-11。

$$A_{hf}^{i} = PRelu\left(F_{MDE}^{i} - F_{blurred}^{i}\right) \tag{3-11}$$

式中，A_{hf}^{i} —— 获得的第 i 层的高频注意力；

$PRelu(\cdot)$ —— 带参数的修正线性单元，即激活函数。

接着，使用获得的高频注意力来对单目深度估计子网络提取的原始特征进行修正和优化，通过残差连接最终得到优化后的引导特征，即式 3-12。

$$F_{hg}^{i} = F_{MDE}^{i} + A_{hf}^{i} \cdot F_{MDE}^{i} \tag{3-12}$$

式中，F_{hg}^{i} —— 第 i 层优化后的引导特征。

定义上述操作的原因是要在单目深度估计的原始特征中突显高频信息，以便低分辨率深度图像可以在特征融合时最大化地利用其中的高频信息。

为了利用来自单目深度估计子网络编码器优化后的指导特征，首先将它们与深度图像超分辨率重建子网络编码器相应层的特征在通道维度级联起来，以生成复合特征 F_{comp}^{i}。这种简单的特征融合在空间维度和通道维度上会有很多冗余，因此本文引入了一个注意力块，其中包括一个通道注意力[25]和一个空间注意力[25]来增强特征融合能力。通道注意力会学习每个特征通道的重要性，而空间注意会突出显示特征图中的重要位置。上述过程可以表述为式 3-13 和式 3-14。

$$F_{comp}^{i} = \left[F_{DSR}^{i}, F_{hg}^{i}\right] \tag{3-13}$$

$$F_{ha}^{i} = SA\left(conv_{1\times1}\left(CA\left(F_{comp}^{i}\right)\right)\right) \tag{3-14}$$

式中，F_{ha}^{i} —— 深度图像超分辨率重建子网络第 i 层融合高频信息的特征；

F_{DSR}^{i} —— 深度图像超分辨率重建子网络第 i 层的特征；

CA —— 通道注意力；

SA —— 空间注意力；

$conv_{1\times1}$ —— 卷积核大小为 1×1 的卷积层；

$[\cdot,\cdot]$ —— 通道维度的级联。

融合了高频信息的特征 F_{ha}^{i} 将作为深度图像超分辨率重建子网络编码器下一层的输入。

本文采用的通道注意力模块，主要借鉴了卷积注意力机制模块（CBAM: Convolutional Block Attention Module）[25]中通道注意力模块的设计，如图 3-6 所示。

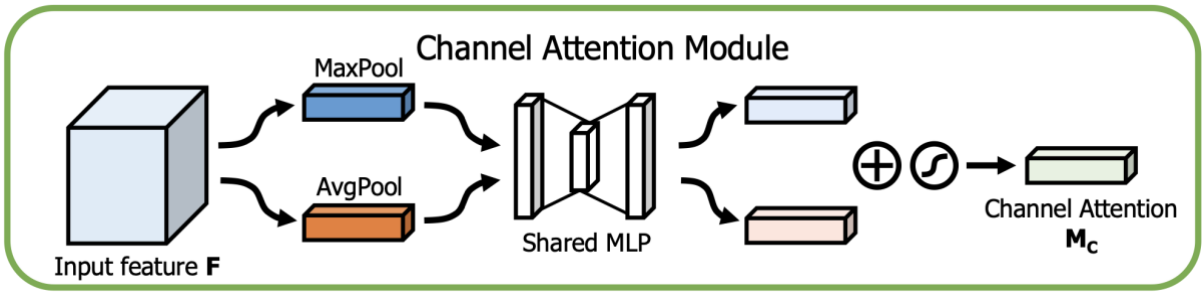

图 3-6 通道注意力模块结构图

对于多通道特征图而言，其每一个通道都可以看作是一个独立的“特征”，因此通道注意力关注的是“对于图像而言什么特征是重要的”。为了降低学习通道注意力时的计算开销，最常用的方法是对特征图的空间维度进行“压缩”，现有的大多数工作通常只采用平均池化来聚合特征图的空间信息（如压缩-激励网络）。除平均池化外，最大池化也可用于聚合特征图所蕴含的空间信息，不同的是在梯度下降时平均池化对特征图上每一个像素点都有反馈，而最大池化只对特征图中响应最大的位置有反馈。实验表明最大池化可以挖掘到图像特征的其他重要“线索”，因而可以与平均池化合作学习得到更加有效的通道注意力。因此，本文同时使用平均池化和最大池化来对特征图的空间信息进行压缩，生成两个不同的上下文描述 F_{avg}^{c} 和 F_{max}^{c}，然后将他们输入到一个共享的网络，并对网络的输出逐像素求和，最后经过 Sigmoid 激活函数以生成通道注意力。整个过程可以描述为式 3-15。

$$\begin{aligned} A^{C}(F) &= \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^{c}\right)\right) + W_1\left(W_0(F_{max}^{c})\right)\right) \end{aligned} \qquad (3-15)$$

式中，$A^{C}(F)$ —— 通道注意力；

σ —— Sigmoid 激活函数；

MLP —— 共享的网络；

$AvgPool$ —— 平均池化操作；

$MaxPool$ —— 最大池化操作；

F —— 特征图；

W_0，W_1 —— 共享网络的权重。

本文采用的空间注意力模块，主要借鉴了卷积注意力机制模块（CBAM: Convolutional Block Attention Module）[25]中空间注意力模块的设计，如图 3-7 所示。空间注意力是基于像素之间的空间关系产生的。与通道注意力不同，空间注意力关注的是“对于图像而言哪些位置是重要的”。类似地，本文采用平均池化和最大池化来对特征

图的通道维度进行压缩，从而生成两个通道数为 1 的特征图 F^s_{avg} 和 F^s_{max}，分别代表原始特征所有通道的平均响应和最大响应。然后将它们级联起来并通过卷积核大小为 7×7 的卷积层，最后经过 Sigmoid 激活函数以生成空间注意力。

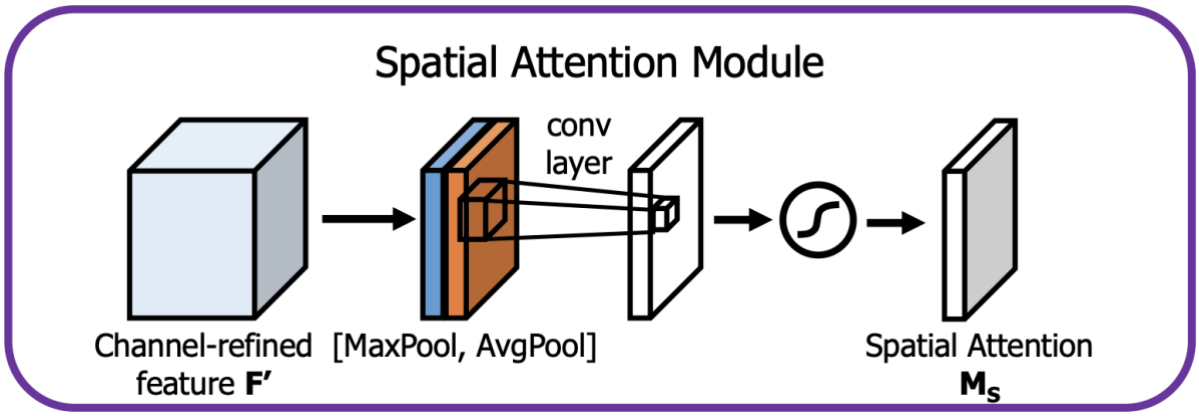

图 3-7 空间注意力模块结构图

整个过程可以描述为式 3-16。

$$A^S(F) = \sigma\left(f^{7\times7}\left([AvgPool(F);\ MaxPool(F)\right)\right)$$
$$= \sigma\left(f^{7\times7}([\ F^s_{avg};\ F^s_{max}])\right) \qquad (3-16)$$

式中，$A^S(F)$ —— 空间注意力；

σ —— Sigmoid 激活函数；

$f^{7\times7}$ —— 卷积核大小为 7×7 的卷积层；

$AvgPool$ —— 平均池化操作；

$MaxPool$ —— 最大池化操作；

F —— 特征图。

3.3 内容引导桥

在特征解码阶段，深度图像超分辨率重建子网络和单目深度估计子网络解码器的作用是进一步提取面向任务的特征，以完成深度估计和深度图像的超分辨率重建，最终可以从两个子网络获得相应的估计或超分辨重建的深度图像。两个任务相较而言，单目深度估计由于尺度模糊性[17]而被广泛认知为不适定的逆问题。例如，世界上观察到的许多三维场景可以对应于完全相同的二维平面，即三维场景与二维平面之间是多对一的关系，如图 3-8 所示。在人眼或相机的成像过程中，大而远的物体和小而近的物体在同一成像平面上的二维信息是一致的，其前提是大物体和小物体具有相同的外观。一个很极端的例子便是现实的房间和其缩小后的展示模型。也就是说在丢失了深度信息后，想要从二

维图像推理出场景的真实深度是有很大难度的，这便是尺度模糊性。

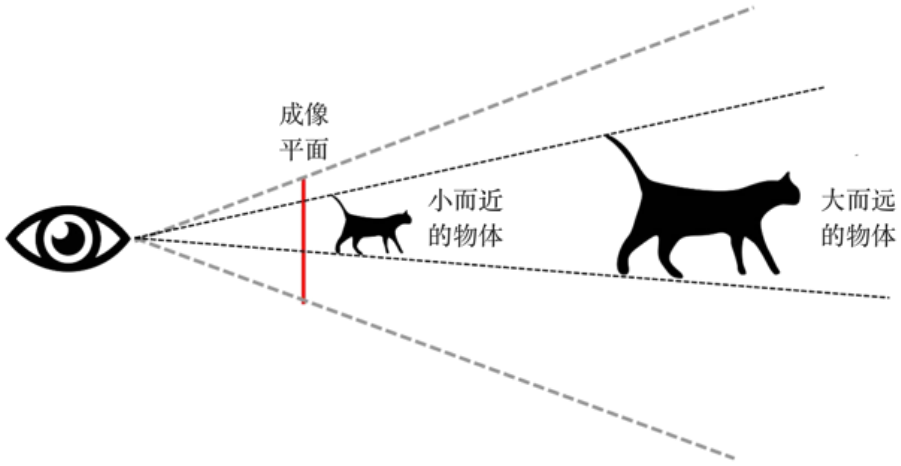

图 3-8 尺度模糊性示例图

因此，训练一个可以很好地从彩色图像映射到深度图像的模型是一项非常艰巨的任务。尽管深度图像超分辨率重建也是一个不适定的问题，但它仍在相同的域中学习映射关系并专注于还原图像的细节，故其相对单目深度估计而言更简单。因此，由于两个任务的性能之间存在较大差距，单目深度估计子网络的解码器生成的特征不再适合为深度图像超分辨率重建的解码器提供指导信息。遵循简单任务指导困难任务的原则，本文在解码阶段交换了两个子网络的指导身份，即让深度图像超分辨率重建子网络在深度特征空间为单目深度估计子网络提供内容引导。详细结构如图 3-9 所示。

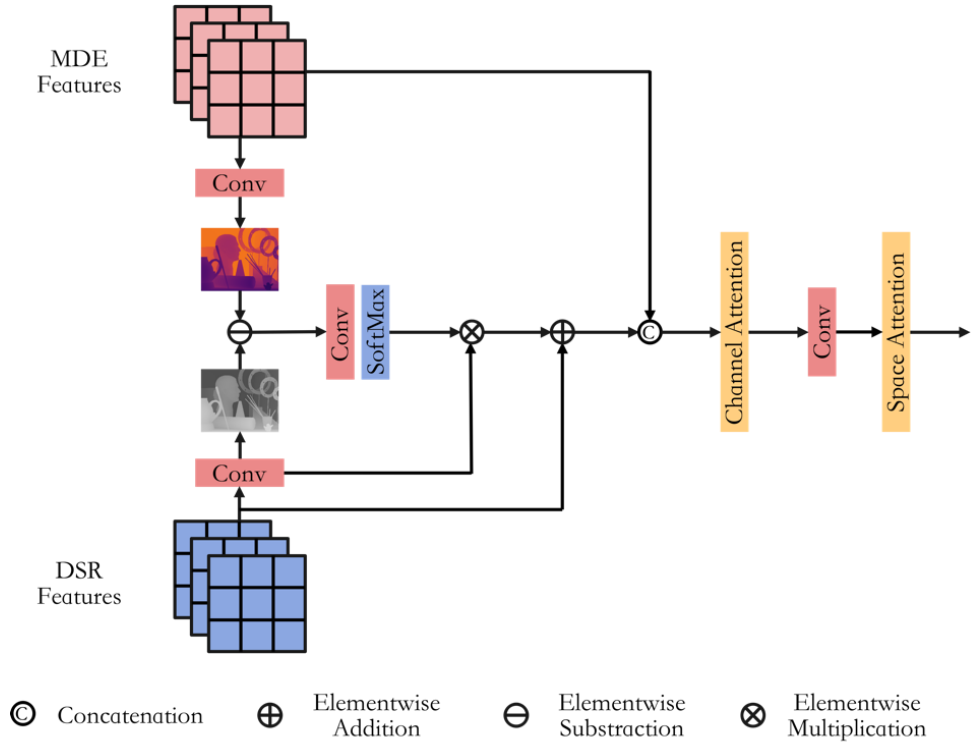

图 3-9 内容引导桥结构图

如前所述，本文提出的模型可以通过两个子网络的解码器特征获得相应的深度图像。具体来说，本文采用卷积核大小为 1×1 的卷积层分别作用于深度图像超分辨率重建子网络和深度估计子网络的解码器，从而获得超分辨率的深度图像和估计的深度图像，即式 3-17 和式 3-18。

$$M_{DSR}^{i}=conv_{1\times1}\left(Fd_{DSR}^{i}\right) \tag{3-17}$$

$$M_{MDE}^{i}=conv_{1\times1}\left(Fd_{MDE}^{i}\right) \tag{3-18}$$

式中，M_{DSR}^{i} —— 深度图像超分辨率重建子网络第 i 层生成的深度图像；

M_{MDE}^{i} —— 单目深度估计子网络第 i 层估计的深度图像；

Fd_{DSR}^{i} —— 深度图像超分辨率重建子网络解码器第 i 层的特征；

Fd_{MDE}^{i} —— 单目深度估计子网络解码器第 i 层的特征。

然后，可以计算得到估计的深度图像 M_{MDE}^{i} 与超分辨率重建的深度图像 M_{DSR}^{i} 之间的差异图。差异图突出显示了估计的深度图像中相对于超分辨重建的深度图像需要进一步优化的位置，并希望这种差异会随着网络的训练越来越小。在此基础上，本文通过对差异图应用卷积运算和 softmax 激活来学习差异权重，从而为单目深度估计子网络提供内容引导。上述操作可以被描述为式 3-19 和式 3-20。

$$W_{diff}^{i}=softmax\left(conv_{1\times1}\left(M_{DSR}^{i}-M_{MDE}^{i}\right)\right) \tag{3-19}$$

$$F_{cg}^{i}=Fd_{DSR}^{i}+W_{diff}^{i}*Fd_{DSR}^{i} \tag{3-20}$$

式中，W_{diff}^{i} —— 差异权重；

F_{cg}^{i} —— 第 i 层的内容引导特征；

$softmax$ —— softmax 激活函数。

最后，本文使用与高频注意力桥中相同的注意力块（包含一个通道注意力和空间注意力）来优化级联的特征（即，$F_{con}^{i}=\left[Fd_{MDE}^{i},F_{cg}^{i}\right]$），优化后的特征将作为单目深度估计子网络解码器中下一层 CLIFF 模块的高层特征输入，并与来自特征金字塔网络的底层特征进一步融合。

3.4 本章小结

本章从模型的组成和联合学习策略的选择，到深度图像超分辨率重建和单目深度估计两个任务间交互模式的设计进行了详细的介绍。本章介绍了深度图像超分辨率重建和单目深度估计的联合学习网络，以实现更好的深度图像超分辨率重建的性能。与现有工作[22]不同的是，在设计两个子网络之间的交互时，本文采用了更为明确的指导模式。在特征编码器中，单目深度估计子网络通过高频注意力桥为深度图像超分辨率重建子网络提供高频信息的指导。与传统的颜色指导的深度图像超分辨率重建算法相比，单目深度估计子网络提供的颜色指导更接近于深度模态。在特征解码器中，本文遵循简单任务指

导困难任务的原则，深度图像超分辨率重建子网络通过内容引导桥为单目深度估计子网络提供内容引导。除了在任务间交互层面的设计外，本章也介绍了在损失函数层级上对两个任务联合优化的策略。

4 实验测试与分析

本章将介绍本文提出的单目深度估计和深度图像超分辨率重建联合学习网络在所选择的公开基准数据集上训练和测试取得的效果。第一小节将介绍为验证本文提出的单目深度估计和深度图像超分辨率重建联合学习网络有效性所用到的数据集及选择的评价指标。第二小节将对数据集的训练集和测试集以及具体的参数设置等进行介绍。第三小节将对本文提出的算法在 Middlebury 数据集上取得的结果进行展示并与现有的深度图像超分辨率重建算法进行比较和分析。第四小节将介绍本文提出的算法在 NYU v2 数据集上训练和测试的结果，同样将与现有算法进行比较分析。最后，第五小节通过设计消融实验验证了本文不同模块设计对整体网络性能的影响。

4.1 数据集及评价指标

4.1.1 数据集

（1） Middlebury

Middlebury 数据集由一系列分别在 2001 年，2005 年，2006 年和 2014 年中引入的子数据集组成。 Middlebury 数据集是在实验室中获取的，其场景仅涵盖不同对象的近景，部分彩色图像和对应的深度图像如图 4-1 所示。

图 4-1 Middlebury 数据集彩色图像和深度图像示例

（2） NUY v2

NYU v2 的原始数据集包含了 464 个室内场景，是由 RGB-D 摄像机获取的场景的单目视频序列和对应的深度真值。其中，1449 个图像对被标注了密集标签且已经对缺失值进行了填充，可以直接用于深度图像超分辨率重建任务。这些带有标签的图像分辨率为 640 × 480，部分彩色图像和对应的深度图像如图 4-2 所示。

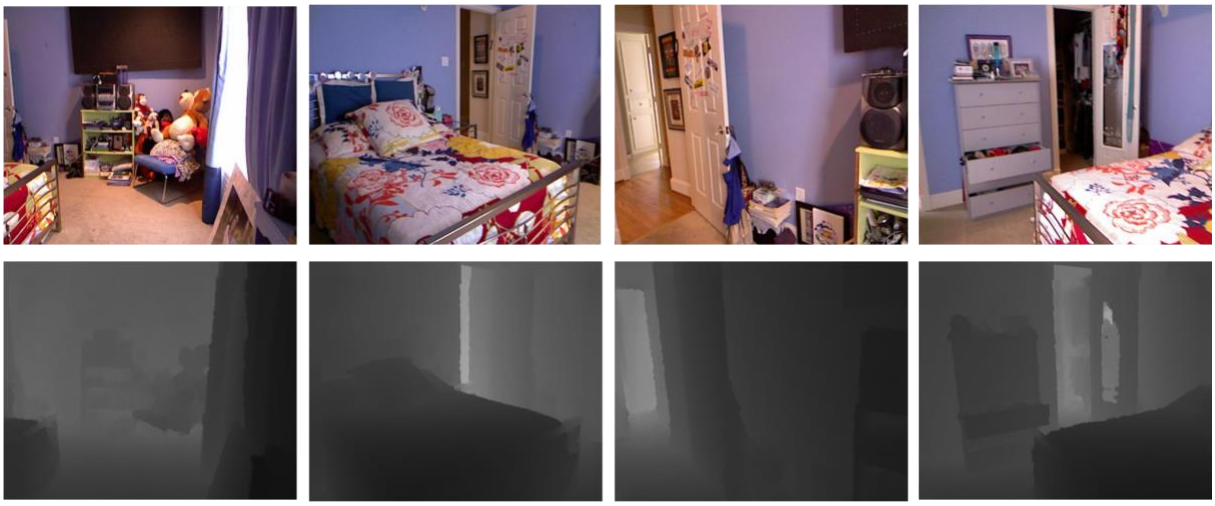

图 4-2 NYU v2 数据集彩色图像和深度图像示例

4.1.2 评价指标

本文选择了平均绝对差（Mean Absolute Difference, MAD）和均方根误差（Root Mean Square Error, RMSE）来度量超分辨率重建得到的深度图像与真值之间的差异。

对于 Middlebury 数据集 $X = \{I_{HR}^n, D_{HR}^n, D_{LR}^n\}_{n=1}^N$，$I_{HR}^n, D_{HR}^n, D_{LR}^n$ 分别为高分辨率彩色图像和深度图像以及对应的低分辨率深度图像，N 为数据集中图像的数量，则平均绝对差可以定义为式 4-1。

$$MAD_{Middlebury} = \frac{1}{N}\sum_{n=1}^{N}|D_{HR}^n - SR(D_{LR}^n)| \qquad (4-1.)$$

式中，SR —— 超分辨率重建操作。

同样，对于 NYU v2 数据集 $Y = \{I_{HR}^m, D_{HR}^m, D_{LR}^m\}_{m=1}^M$，$I_{HR}^m, D_{HR}^m, D_{LR}^m$ 分别为高分辨率彩色图像和深度图像以及对应的低分辨率深度图像，M 为数据集中图像的数量，则均方根误差可以定义为式 4-2。

$$RMSE_{NYU\ v2} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(D_{HR}^m - SR(D_{LR}^m)\right)^2} \qquad (4-2.)$$

式中，SR —— 超分辨率重建操作。

4.2 实验配置

本文从 Middlebury 数据集中收集了 36 对 RGB-D 图像对（分别来自 2001[35]，2006[36]和 2014[37]的 6、21、9 张图像）用于训练，将 Middlebury 2005[38]数据集中的 6 对 RGB-D 对图像对（分别为 Art，Books，Moebius，Dolls，Laundry，Reindeer）用于测试。本文选取的另一个用于训练和测试的数据集是 NYU v2 数据集[39]，并按照通用的划分方法[40]，将数据集中的前 1000 对 RGB-D 图像对作为训练数据，然后在后 449 对 RGB-D 图像对上进行评估。此外，用于训练和测试的 RGB-D 图像对被归一在 [0，1] 范围内。

遵循现有工作的方法[10]，本文在高分辨率图像输入网络前将其划分为具有重叠的规则网格，即把一幅完整的图像裁剪成足够数量的图像块。这种训练策略不仅不会削弱网络的性能，而且会减少训练的时间。高分辨率的深度图像和彩色图像分别根据 $\times 4$、$\times 8$ 和 $\times 16$ 的上采样因子被裁剪成足够数量（在本文实验中一般裁剪为 15000 张左右的图像块）的大小为 64^2、128^2 和256^2 的图像块。为了获得相应的低分辨率深度图像块，本文使用 Bicubic 插值方法将高分辨率深度图像块下采样为固定大小的 16×16 的图像块。如前文所述，本文引入了平均绝对差（MAD）和均方根误差（RMSE）两个指标，以对模型的性能进行定量评估。

本文提出的网络基于深度学习框架 PyTorch 实现，并使用 NVIDIA 2080Ti GPU 加速训练。在训练期间，一次训练所选取的样本数（Batch Size）为 8。此外，本文选取了动量为 0.9，$\beta_1 = 0.9$，$\beta_2 = 0.99$，$\epsilon = 10^{-8}$ 的 ADAM 优化器对训练进行优化。初始学习率设置为 $1e^{-4}$，并且每 100 轮（epoch）乘以 0.1 以降低学习速率。在上采样因子为 $\times 8$ 的深度图像超分辨率重建实验中，大小为 256×256 的图像的推理时间为 0.052 秒。

4.3 基于 Middlebury 数据集的结果比较及分析

本文在不同上采样因子（$\times 4$，$\times 8$ 和 $\times 16$）与一些最新的深度图像超分辨率重建方法进行了比较，包括六种传统深度图像超分辨率重建算法（即 CLMF[7]，JGF[6]，TGV[8]，CDLLC[41]，PB[42] 和 EG[43]）和七种基于深度学习的方法（即 SRCNN[44]，ATGVNet[45]，MSG[10]，DGDIE[46]，DEIN[34]，CCFN[11]，GSRPT[12] 和 CTKT[21]）。

图 4-3 展示了 $\times 8$ 上采样因子下不同方法在图像 Art 和 Dolls 上的可视化结果比较。其中，（a）为深度图像的真值（Ground Truth）和彩色图像；（b）为低分辨率深度块；（c）-（h）分别为通过 Bicubic，TGV[8]，MSG[10]，CTKT[21] 和 BridgeNet 超分辨率重建得到的深度图像块；（i）为深度图像的真值图像块。为了获得更加清晰的可视化

结果，深度图像块经过了放大。

显然，就场景的结构和细节而言，与深度图像的真值相比，本文的方法获得了最相似的重建结果。对于图中的尺寸较大的物体（如 Art 中的茶壶），所有的基于深度学习的方法都呈现出相似的重建结果。然而，对于微小的物体，本文的方法可以恢复深度图像更多的细节。例如，*Art* 图像中的棍子周围的伪影更少，*Doll* 图像中的玩具头部轮廓更准确。而其他方法可能会产生一些伪像，边界模糊或形状变化。这些现象与低分辨率深度图像成像时深度相机对精细结构和微小物体的严重破坏有关，从而给这些区域的重建带来了更多困难。因此，本文的方法在准确恢复这些微小物体的深度边界方面更具有优势。

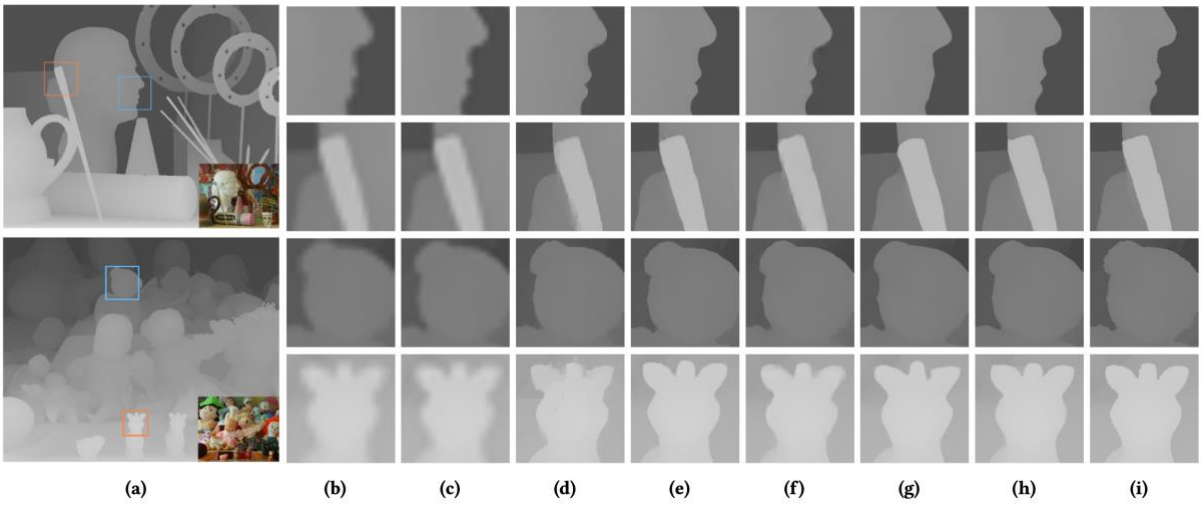

图 4-3 Middlebury 数据集 ×8 不同超分辨率重建方法可视化结果对比

与其他方法的定量比较如表 4-1 和表 4-2 所示（性能最好的模型指标被加粗显示，而性能排名第二的模型指标被下划线标记）。与基于卷积神经网络的方法相比，传统的基于滤波或基于优化的方法具有相对较高的平均绝对差值。而与其他基于卷积神经网络的方法相比，本文提出的单目深度估计和深度图像超分辨率重建联合学习网络（BridgeNet）可以获得具有竞争力的结果，即使在 ×8 和 ×16 这样具有挑战性的上采样因子下，本文提出的网络也几乎可以获得最佳的结果。以上采样因子为 ×16 的深度图像超分辨率重建实验为例，对于 Books 图像，与性能排名第二的算法相比，本文的方法将 MAD 从 0.67 提升至 0.51，百分比增益为 23.9%。

表 4-1 Middlebury 数据集 ×8 不同超分辨率重建方法量化指标对比（一）

	Art			Books			Dolls		
	×4	×8	×16	×4	×8	×16	×4	×8	×16
CLMF[7]	0.76	1.44	2.87	0.28	0.51	1.02	0.34	0.60	1.01
JGF[6]	0.47	0.78	1.54	0.24	0.43	0.81	0.33	0.59	1.06

表 4-1（续）

	Art			Books			Dolls		
	× 4	× 8	× 16	× 4	× 8	× 16	× 4	× 8	× 16
TGV[8]	0.65	1.17	2.30	0.27	0.43	0.82	0.33	0.70	2.20
CDLLC[41]	0.53	0.76	1.41	0.19	0.46	0.75	0.31	0.53	0.79
PB[42]	0.79	0.93	1.98	0.16	0.43	0.79	0.53	0.83	0.99
EG[43]	0.48	0.71	1.35	0.15	0.36	0.70	0.27	0.49	0.74
SRCNN[44]	0.63	1.21	2.34	0.25	0.52	0.97	0.29	0.58	1.03
ATGVNet[45]	0.65	0.81	1.42	0.43	0.51	0.79	0.41	0.52	**0.56**
MSG[10]	0.46	0.76	1.53	0.15	0.41	0.76	0.25	0.51	0.87
DGDIE[46]	0.48	1.20	2.44	0.30	0.58	1.02	0.34	0.63	0.93
DEIN33]	0.40	0.64	**1.34**	0.22	0.37	0.78	0.22	0.38	0.73
CCFN[11]	0.43	0.72	1.50	0.17	0.36	0.69	0.25	0.46	0.75
GSRPT[12]	0.48	0.74	1.48	0.21	0.38	0.76	0.28	0.48	0.79
CTKT[21]	**0.25**	**0.53**	1.44	**0.11**	0.26	0.67	**0.16**	0.36	0.65
BridgeNet	0.30	0.58	1.49	0.14	**0.24**	**0.51**	0.19	**0.34**	0.64

表 4-2 Middlebury 数据集 × 8 不同超分辨率重建方法量化指标对比（二）

	Laundry			Mobius			Reindeer		
	× 4	× 8	× 16	× 4	× 8	× 16	× 4	× 8	× 16
CLMF[7]	0.50	0.80	1.67	0.29	0.51	0.97	0.51	0.84	1.55
JGF[6]	0.36	0.64	1.20	0.25	0.46	0.80	0.38	0.64	1.09
TGV[8]	0.55	1.22	3.37	0.29	0.49	0.90	0.49	1.03	3.05
CDLLC[41]	0.30	0.48	0.96	0.27	0.46	0.79	0.43	0.55	0.98
PB[42]	1.13	1.89	2.87	0.17	0.47	0.82	0.56	0.97	1.89
EG[43]	0.28	0.45	0.92	0.23	0.42	0.75	0.36	0.51	0.95
SRCNN[44]	0.40	0.87	1.74	0.25	0.43	0.87	0.35	0.75	1.47
ATGVNet[45]	0.37	0.89	0.94	0.38	0.45	0.80	0.41	0.58	1.01
MSG[10]	0.30	0.46	1.12	0.21	0.43	0.76	0.31	0.52	0.99
DGDIE[46]	0.35	0.86	1.56	0.28	0.58	0.98	0.35	0.73	1.29
DEIN33]	0.23	0.36	0.81	0.20	0.35	0.73	0.26	0.40	0.80
CCFN[11]	0.24	0.41	0.71	0.23	0.39	0.73	0.29	0.46	0.95
GSRPT[12]	0.33	0.56	1.24	0.24	0.49	0.80	0.31	0.61	1.07

表 4-2（续）

	Laundry			Mobius			Reindeer		
	×4	×8	×16	×4	×8	×16	×4	×8	×16
CTKT[21]	**0.16**	0.36	0.76	**0.13**	0.27	0.69	**0.17**	0.35	0.77
BridgeNet	0.17	**0.34**	**0.71**	0.15	**0.26**	**0.54**	0.19	**0.31**	**0.70**

4.4 基于 NYU v2 数据集的结果比较及分析

本文还在 NYU v2 数据集上对提出的方法进行评估，并与其他最新的方法进行比较，包括 Bicubic，TGV[8]，EDGE[9]，DJF[40]，SDF[47]，DGDIE[46]，GbFT[48]，PAC[49]，SVLRM[50]，DKN[51] 和 CTKT[21]。

图 4-4 展示了本文的方法在 ×8 上采样因子下的可视化结果。其中，（a）为彩色图像；（b）为深度图像真值；（c）-（f）分别为 SDF[47]，DJF[40]，SVLRM[50] 和 BridgeNet 超分辨率重建得到的深度图像块。可以看到，无论是在红色矩形区域还是在黄色矩形区域内，本文的方法都可以准确地重建图像的深度信息和微小物体的边缘。如表 4-3 中所示（性能最好的模型指标被加粗显示，而性能排名第二的模型指标被下划线标记），在 ×8 和 ×16 上采样因子的情况下，本文的方法可以获得最佳的性能。与性能排名第二的算法相比，本文方法的 RMSE 在 ×8 的上采样因子下达到 2.63，提升了 3.6%。

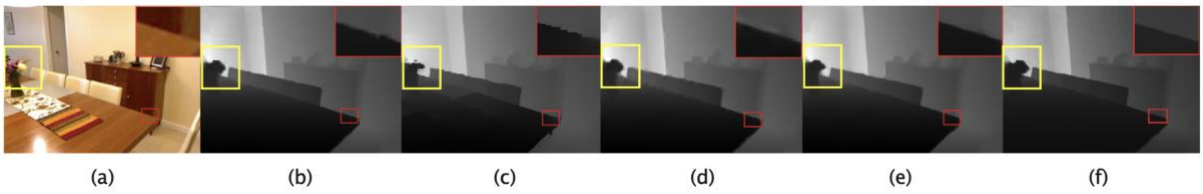

图 4-4 NYU v2 数据集 ×8 不同超分辨率重建方法可视化结果对比

表 4-3 NYU v2 数据集 ×8 不同超分辨率重建方法量化指标对比

	×4	×8	×16
Bicubic	8.16	14.22	22.32
TGV[8]	6.98	11.23	28.13
EDGE[9]	5.21	9.56	18.1
DJF[40]	3.54	6.2	10.21
SDF[47]	3.04	5.67	9.97
DGDIE[46]	1.56	2.99	5.24
GbFT[48]	3.35	5.73	9.01
PAC[49]	2.39	4.59	8.09

表 4-3（续）

	×4	×8	×16
SVLRM[50]	1.74	5.59	7.23
DKN[51]	1.62	3.26	6.51
CTKT[21]	**1.49**	2.73	5.11
BridgeNet	1.54	**2.63**	**4.98**

4.5 消融实验

在本节中，将进行全面的消融研究以验证单目深度估计和深度图像超分辨率重建联合学习网络（BridgeNet）中的设计对提升深度图像超分辨率重建性能的作用及贡献。表 4-4 列出了 Middlebury 2005 数据集在不同实验设计下进行八倍上采样深度图像超分辨率重建的结果。第一行和第二行是在相同条件下分别单独对深度图像超分辨率重建子网络和单目深度估计子网络训练与测试的结果。从平均绝对差的值来看，单目深度估计子网络的性能不及深度图像超分辨率重建子网络，这样的结果与之前的分析吻合。也就是说，单目深度估计任务的学习难度与深度图像超分辨率重建任务相比较为困难。

对于任务间的交互，本文首先通过简单的损失函数约束将两个任务组合在一起，结果为表 4-4 的第三行。与单独的深度图像超分辨率重建子网络相比，深度图像超分辨率重建的 MAD 优化到 0.363。然后，逐渐将高频注意力桥（HABdg）和内容引导桥（CGBdg）集成到网络中以验证它们的作用。在联合学习网络中仅使用高频注意力桥或内容引导桥时，获得的结果要好于仅使用损失函数约束。此外，当两个桥接器一起工作时（即完整模型），网络的性能将达到最佳。本文还在图 4-5 中提供了一些可视化的比较，其中（a）为深度图像真值；（b）为单独的深度图像超分辨率重建子网络（DSRNet）的结果；（c）为 BridgeNet 的结果。从图中可以看出，与单独的深度图像超分辨率重建子网络相比，本文的模型具有更清晰的边界和更准确的深度值，如图中的黄色矩形区域所示。

表 4-4 针对 BridgeNet 的消融研究量化对比

	DSRNet	MDENet	HABdg	CGBdg	Middlebury
1	√				0.366
2		√			0.472
3	√	√			0.363
4	√	√	√		0.355
5	√	√		√	0.361
6	√	√	√	√	0.343

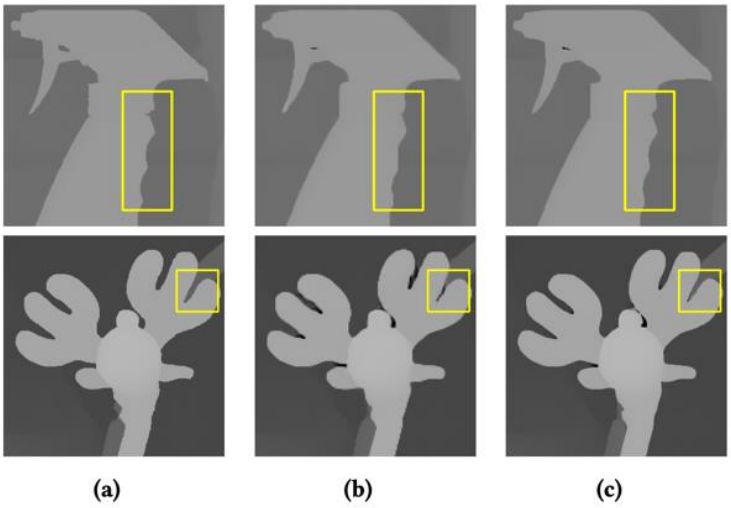

图 4-5 消融实验可视化结果对比

为了进一步证明高频注意力桥的有效性，本文通过将单目深度估计子网络的特征未经任何处理地馈入到深度图像超分辨率重建子网络来代替它，并在 Middlebury 数据集上进行了训练，以及在 Middlebury 2005 数据集上进行了上采样因子为 $\times 8$ 的对比实验，量化指标结果如表 4-5 所示（w/o HABdg 指的是通过直接将单目深度估计子网络的特征传递给深度图像超分辨率重建子网络代替高频注意力桥）。与简单地将通道维度上的特征级联并送入深度图像超分辨率重建网络相比，本文提出的高频注意力桥可以提供更有效的高频指导以提高性能。

表 4-5 针对高频注意力桥的消融研究量化对比

	Art	Books	Dolls	Laundry	Mobius	Reindeer	Avg.
w/ HABdg	0.58	0.24	0.34	0.34	0.26	0.31	0.343
w/o HABdg	0.65	0.25	0.37	0.38	0.28	0.33	0.376

4.6 本章小结

本章主要介绍了本文所设计的实验及其测试结果，包括在不同数据集上与最新方法的可视化对比和评价指标的比较。首先，介绍了本文采用的公开基准数据集和评估指标，并对有关实验的详细配置进行了阐明，然后介绍了本文提出的单目深度估计和深度图像超分辨率重建联合学习网络在 Middlebury 数据集上不同上采样因子的实验结果及分析，并与最新的深度超分辨率重建算法的结果进行了对比。同样，本文也将 NYU v2 数据集上不同上采样因子的实验结果与代表性的方法进行了对比和分析。此外，本文通过设计详尽的消融实验探究了所提出网络设计的有效性，验证了两个桥接器对提升深度图像超分辨率重建任务性能的积极作用。

5 结论

本文探索了一种联合学习框架，该框架结合了深度图像超分辨率重建和单目深度估计两个任务，可以在不添加任何其他监督信息的情况下提升深度图像超分辨率重建的性能。

高分辨率彩色图像由于其具有与深度图像的结构相似性且容易获得，被广泛用于为深度图像超分辨率重建提供先验信息。但现有颜色指导的深度图像超分辨率重建算法通过额外分支提取到的指导信息，并没有很好地面向深度模态，因此可能在两种模态结构不一致的区域造成纹理复制等问题。而单目深度估计旨在将场景从光度表示映射到几何表示，换言之，在连续的训练和学习过程中，单目深度估计实现了从彩色图像到深度图像的跨模态信息转换。因而面向单目深度估计学习到的彩色图像特征更适合指导深度图像超分辨率重建。

本文的核心思想是如何设计两个子网络（即深度图像超分辨率重建子网络和单目深度估计子网络）之间的交互，由此本文提出了两个桥接器。特征编码阶段中的高频注意力桥将从单目深度估计子网络学习到的彩色高频信息传输到深度图像超分辨率重建子网络，从而可以提供更接近深度模态的颜色指导信息。遵循简单任务指导困难任务的原则，在特征解码阶段交换了两个任务的指导角色，深度图像超分辨率重建子网络通过内容引导桥为单目深度估计子网络在深度特征空间提供内容引导。全面的实验表明，本文提出的方法达到了具有竞争力的性能，尤其是在上采样因子较大的情况下。

此外，本文提出的网络结构具有高度的可移植性，可以为关联深度图像超分辨率重建任务和单目深度估计任务提供范例。在未来的工作中，可以通过替换不同的单目深度估计子网络和深度图像超分辨率重建子网络以更进一步验证本文提出的交互模式的普适性，并在提升深度图像超分率重建性能的同时加快网络的推理速度。

参考文献

[1] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras[C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013: 2100-2106.

[2] Im S, Ha H, Choe G, et al. Accurate 3d reconstruction from small motion clip for rolling shutter cameras[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(4): 775-787.

[3] Fan D P, Lin Z, Zhang Z, et al. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks[J]. IEEE Transactions on neural networks and learning systems, 2020.

[4] Ge L, Liang H, Yuan J, et al. Real-time 3D hand pose estimation with 3D convolutional neural networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(4): 956-970.

[5] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 746-760.

[6] Liu M Y, Tuzel O, Taguchi Y. Joint geodesic upsampling of depth images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 169-176.

[7] Lu J, Shi K, Min D, et al. Cross-based local multipoint filtering[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 430-437.

[8] Ferstl D, Reinbacher C, Ranftl R, et al. Image guided depth upsampling using anisotropic total generalized variation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 993-1000.

[9] Park J, Kim H, Tai Y W, et al. High quality depth map upsampling for 3d-tof cameras[C]//2011 International Conference on Computer Vision. IEEE, 2011: 1623-1630.

[10] Hui T W, Loy C C, Tang X. Depth map super-resolution by deep multi-scale guidance[C]//European conference on computer vision. Springer, Cham, 2016: 353-369.

[11] Wen Y, Sheng B, Li P, et al. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution[J]. IEEE Transactions on Image Processing, 2018, 28(2): 994-1006.

[12] Lutio R, D'aronco S, Wegner J D, et al. Guided super-resolution as pixel-to-pixel transformation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8829-8837.

[13] Zhao L, Bai H, Liang J, et al. Simultaneous color-depth super-resolution with conditional generative adversarial networks[J]. Pattern Recognition, 2019, 88: 356-369.

[14] Zuo Y, Fang Y, Yang Y, et al. Residual dense network for intensity-guided depth map enhancement[J]. Information Sciences, 2019, 495: 52-64.

[15] Ye X, Sun B, Wang Z, et al. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution[J]. IEEE Transactions on Image Processing, 2020, 29: 7427-7442.

[16] Guo C, Li C, Guo J, et al. Hierarchical features driven residual learning for depth map super-resolution[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2545-2557.

[17] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[J]. arXiv preprint arXiv:1406.2283, 2014.

[18] 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.

[19] Cao Y, Wu Z, Shen C. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(11): 3174-3182.

[20] Zhang Z, Cui Z, Xu C, et al. Joint task-recursive learning for semantic segmentation and depth estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 235-251.

[21] He L, Lu J, Wang G, et al. SOSD-Net: Joint semantic object segmentation and depth estimation from monocular images[J]. Neurocomputing, 2021, 440: 251-263.

[22] Sun B, Ye X, Li B, et al. Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution[J]. arXiv preprint arXiv:2103.12955, 2021.

[23] Qiu X. Neural Networks and Deep Learning[J]. 2020.

[24] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European conference on computer vision. Springer, Cham, 2016: 483-499.

[25] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[26] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.
[27] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[28] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[J]. arXiv preprint arXiv:1506.02025, 2015.
[29] Wang L, Zhang J, Wang Y, et al. CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss[C]//European Conference on Computer Vision. Springer, Cham, 2020: 316-331.
[30] Yin Y, Robinson J, Zhang Y, et al. Joint super-resolution and alignment of tiny faces[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12693-12700.
[31] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 136-144.
[32] Zhang Y, Tian Y, Kong Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2472-2481.
[33] Wang J, Xu W, Cai J F, et al. Multi-Direction Dictionary Learning Based Depth Map Super-Resolution With Autoregressive Modeling[J]. IEEE Transactions on Multimedia, 2020, 22(6):1470-1484.
[34] Ye X, Duan X, Li H. Depth super-resolution with deep edge-inference network and edge-guided depth filling[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 1398-1402.
[35] Baker S, Scharstein D, Lewis J P, et al. A database and evaluation methodology for optical flow[J]. International journal of computer vision, 2011, 92(1): 1-31.
[36] Hirschmuller H, Scharstein D. Evaluation of cost functions for stereo matching[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-8.
[37] Scharstein D, Hirschmüller H, Kitajima Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth[C]//German conference on pattern recognition. Springer, Cham, 2014: 31-42.
[38] Scharstein D, Pal C. Learning conditional random fields for stereo[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-8.
[39] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 746-760.
[40] Li Y, Huang J B, Ahuja N, et al. Deep joint image filtering[C]//European Conference on Computer Vision. Springer, Cham, 2016: 154-169.
[41] Xie J, Chou C C, Feris R, et al. Single depth image super resolution and denoising via coupled dictionary learning with local constraints and shock filtering[C]//2014 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2014: 1-6.
[42] Mac Aodha O, Campbell N D F, Nair A, et al. Patch based synthesis for single depth image super-resolution[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 71-84.
[43] Xie J, Feris R S, Sun M T. Edge-guided single depth image super resolution[J]. IEEE Transactions on Image Processing, 2015, 25(1): 428-438.
[44] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution[C]//European conference on computer vision. Springer, Cham, 2014: 184-199.
[45] Riegler G, Rüther M, Bischof H. Atgv-net: Accurate depth super-resolution[C]//European conference on computer vision. Springer, Cham, 2016: 268-284.
[46] Gu S, Zuo W, Guo S, et al. Learning dynamic guidance for depth image enhancement[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3769-3778.
[47] Ham B, Cho M, Ponce J. Robust guided image filtering using nonconvex potentials[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(1): 192-207.
[48] AlBahar B, Huang J B. Guided image-to-image translation with bi-directional feature transformation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9016-9025.
[49] Su H, Jampani V, Sun D, et al. Pixel-adaptive convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11166-11175.
[50] Pan J, Dong J, Ren J S, et al. Spatially variant linear representation models for joint

filtering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1702-1711.

[51] Kim B, Ponce J, Ham B. Deformable kernel networks for guided depth map upsampling[J]. arXiv preprint arXiv:1903.11286, 2019.

致　　谢

故事讲了很长，感谢你一直读到这里。但故事还没有结束，写下去才知道梦有多长。

桃李不言，下自成蹊。感谢指导教师冯凤娟老师在我完成毕业论文过程中给予的悉心指导和细致审阅，让我能把这份工作更好地呈现在论文中。从第一次在老师的课堂上学习到完成毕业设计的选题和论文的撰写，我不仅从中获得了很多专业技能的提升，冯老师为人的热情和执教的认真也深深影响了我。谨在此以平铺直叙的话语感谢老师一直以来的教导和帮助，祝愿老师身体健康，工作顺利。

落实思树，饮流怀源。感谢北京交通大学数字媒体信息处理研究中心赵耀老师和丛润民老师提供的研究课题和平台，以及对我研究思路和论文写作的指导。两位老师对科研的热忱更加激发了我对科研的神往和为之努力的斗志。在今后以梦为马的求学路上，我必将紧随两位老师的步伐，把热爱科研的种子埋在心中。同时，也要感谢实验室帮助我顺利完成研究课题的各位同学。

焉得谖草，言树之背。感谢家人一直以来的支持和陪伴，尤其是父母的教育理念培育了我良好的价值观、人生观和世界观。在停课不停学的日子里，我有机会再次和父母放飞了高三曾被放飞的风筝。跨越三年，它承载了我的梦想，也见证了我们一家的幸福时光。父母教会的我不仅是要好好学习，更要热爱自己的生活和这个世界。祝愿我的家人平安喜乐，幸福安康。

学贵得师，亦贵得友。感谢在红果园里每个陪我走过一段路的良师益友，在和同学们相处的难忘岁月里，他们的优良品质影响我颇深。特别感谢在毕设期间陪在身边的三两挚友，我们相聚一起为梦想而努力的欢乐时光必将在未来熠熠闪耀。同时，也要感谢和我一起完成大学生创新训练计划项目的两位朋友，他们是我起航科研星空和叩开计算机视觉大门的“梦想合伙人”。

阳和启蛰，天雨流芳。感谢一路走到现在的自己，或许神明不佑，星辰晦暗，但少年在，光和救赎就在。别想太多，好好生活，日子过着过着就会有答案，努力走着就会有温柔的着落。愿历经千帆，仍记得转身微笑鞠躬，向青春致敬。

喜欢高德地图的一句话：虽然前方拥堵，但您仍在最优路线上。虽然前路艰难，但我相信最好的就快要发生。我期待着明天，前程似锦。

附　录

附录 A 外文文献及翻译

外文文献

Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution

Baoli Sun[1], Xinchen Ye[1,2*], Baopu Li[3], Haojie Li[1,2], Zhihui Wang[1,2], Rui Xu[1,2]

[1]International School of Information Science & Engineering, Dalian University of Technology, China

[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China

[3]Baidu Research, USA

Abstract. Existing color-guided depth super-resolution (DSR) approaches require paired RGB-D data as training samples where the RGB image is used as structural guidance to recover the degraded depth map due to their geometrical similarity. However, the paired data may be limited or expensive to be collected in actual testing environment. Therefore, we explore for the first time to learn the cross-modality knowledge at training stage, where both RGB and depth modalities are available, but test on the target dataset, where only single depth modality exists. Our key idea is to distill the knowledge of scene structural guidance from RGB modality to the single DSR task without changing its network architecture. Specifically, we construct an auxiliary depth estimation (DE) task that takes an RGB image as input to estimate a depth map, and train both DSR task and DE task collaboratively to boost the performance of DSR. Upon this, a cross-task interaction module is proposed to realize bilateral cross-task knowledge transfer. First, we design a cross-task distillation scheme that encourages DSR and DE networks to learn from each other in a teacher-student role-exchanging fashion. Then, we advance a structure prediction (SP) task that provides extra structure regularization to help both DSR and DE networks learn more informative structure representations for depth recovery. Extensive experiments demonstrate that our scheme achieves superior performance in comparison with other DSR methods. Our code available at: https://github.com/Sunbaoli/dsr-distillation.

1. Introduction

To better understand a scene image, depth information is supplemented to the RGB images, providing the key clue about the scene and enabling wide applications in 3D reconstruction, autonomous navigation, monitoring, and so on. However, acquiring depth information for indoor and outdoor scenes needs expensive cost and great efforts, especially for high-quality and high-resolution (HR) depth maps. As such, one of the effective post processing techniques, Depth Super-Resolution (DSR), is greatly desired to yield HR depth maps to alleviate this problem. Many efforts have been taken along the direction of DSR. Usually, fine scene structures are easily lost or severely destroyed in low-resolution (LR) depth map because of the limited spatial resolution. An RGB image and its associated depth map are the photometric and geometrical representations of the same scene, and have a strong structural similarity. Most existing DSR methods learn structural complementarity from RGB images to recover the degraded depth maps.

Previous color-guided DSR methods take advantage of RGB-D image pairs via a two-way fusion architecture, in which an extra branch is required to extract structural guidance from RGB image. As illustrated in Figure 1 (a), RGB image and LR depth map are often processed by separate branches and filtered together through a joint branch to output the HR result. However, due to the simple feature aggregation at a specific layer in the middle of the network, high-frequency structure information from RGB image is more likely to be lost in the process of feature extraction. Therefore, as shown in Figure 1 (b), some novel methods incorporate a new paradigm of feature aggregation, i.e, multi-scale fusion, to allow the network to learn rich hierarchical features at different levels. This in turn makes the network to retain more spatial details for recovering both fine-scale and large-scale structures.

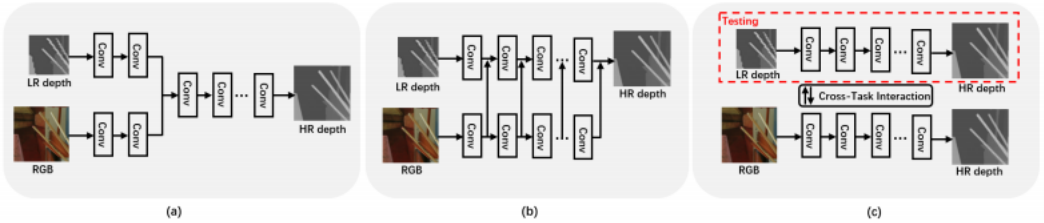

Figure 1. Color-guided DSR paradigms. (a) Joint filtering, (b) Multi-scale feature aggregation, (c) Our cross-task interaction mechanism to distill knowledge from RGB image to DSR task without changing its network architecture.

Although existing color-guided DSR methods have demonstrated remarkable progress, several limitations still remain. First, these methods require paired RGB-D data as training examples to jointly recover the degraded depth map. However, the paired data may be limited or expensive to be collected in actual testing environment. For example, RGB image and depth map are captured by separate depth and RGB sensors with different resolutions and views, thus needing accurate calibration and rectification between them to obtain the registered pairs. Actually, most of real-world applications still come with only a single LR depth

map, which raises the above question. Second, considering the memory consumption and computing burden, the processing on the HR RGB data also hinders the practical application. Moreover, although RGB features can be used as structural guidance to resolve the degradation in DSR, RGB discontinuities do not always coincide with those of depth map (structure inconsistency), which results in noticeable artifacts such as texture copying and depth bleeding. Therefore, how to leverage RGB information to help recover the depth map and simultaneously satisfy the actual testing environment, still needs to be studied.

Motivated by the above analysis, this paper breaks away from the shackles of general paradigms and introduces a novel scene structure guidance learning method for the task of DSR, as shown in Figure 1 (c). We explore for the first time to learn the cross-modality knowledge at training stage, where both RGB and depth modalities are available, but test on the target dataset, where only single depth modality exists. Our key idea is to distill the knowledge of scene structural guidance from RGB modality to the single DSR task without changing its network architecture.

Specifically, as illustrated in Figure 2, inspired by the success of multi-task learning, we construct an auxiliary depth estimation (DE) task that takes RGB image as input to estimate a depth map. Upon this, we propose a cross-task interaction module to realize bilateral knowledge transfer between DSR task and DE task. Different from the commonly used distillation techniques, we first design a cross-task distillation that encourages DSR network (DSRNet) and DE network (DENet) to learn from each other, i.e., the roles of teacher and student will dynamically switch between both tasks based on their current performances on depth recovery in the iterative collaborative training. A multi-space distillation scheme is introduced to distill knowledge from the perspective of output and affinity spaces, which can better describe the essential structural characteristics of depth map. Moreover, to address the problem of RGB-D structure inconsistency, we construct a structure prediction (SP) task that provides extra structure regularization to help both DSRNet and DENet learn more informative structure representations for depth recovery. We come up with an uncertainty-induced attention fusion module to provide a reasonable input for the SP network (SPNet), in which the uncertainty maps acquired from both DSRNet and DENet are used to re-weight their features for strengthening effective structural knowledge. Extensive experiments demonstrate that our single DSR method even outperforms the color-guided DSR methods on benchmark datasets in terms of both accuracy and runtime. The main contributions are summarized as follows,

- So far as we know, our proposed paradigm of DSR is the first work that learns with multiple modalities as inputs at training stage, but tests on only single LR depth modality.
- A cross-task distillation scheme is proposed to encourage DSRNet and DENet to learn from each other in a collaborative training mode.
- A structure prediction network is advanced to provide structure regularization for helping DSRNet resolve the problem of structural inconsistency.

2. Related Work

Depth Super-Resolution. Compared to single DSR methods, color-guided DSR methods have been widely proposed to improve the quality of depth map by the guidance of color image. Li et al. proposed a joint filtering approach that leverages color image as guidance to enhance the spatial resolution of depth map. Hui et al. employed a multi-scale fusion strategy that fuses the rich hierarchical color features at different levels to resolve ambiguity in DSR. Wen et al. presented a data-driven filter to infer an initial HR depth map with the guidance of color image, then proposed a coarse-to-fine network to progressively recover the depth map. Guo et al. proposed a hierarchical feature driven method that constructs an input pyramid and a guidance pyramid for multi-level residual learning. Wang et al. proposed to upsample the depth map with the help of edge map learned from the color image.

Monocular Depth Estimation. Due to the strong ability of CNN in feature extraction, many supervised monocular depth estimation methods continue to improve the performance of depth estimation. Laina et al. proposed a fully convolutional architecture to model the mapping between color image and depth map. In [25], a two-stream CNN is proposed to simultaneously predict depth and depth gradients for accurate depth estimation. Wang et al. presented a depth estimation network via a semantic divide-and-conquer strategy, in which a scene is decomposed into semantic segments and then predicts depth for each segment. In contrast, unsupervised methods use video or stereo data during training without the need of ground truth depth maps. Wong et al. learned a robust representation with a two-branch decoder to estimate the depth map.

Knowledge Distillation. Knowledge distillation is to transfer knowledge from high-capacity model to a compact model to improve the performance of the latter one. It has been wildly applied to many applications, including action recognition, style transfer, depth estimation and scene parsing. For example, the task of image classification takes class probabilities from teacher network as soft targets to train the student network or transfers the knowledge through intermediate layers. Recently, deep mutual learning proposed a two-way distillation which transfers knowledge between the teacher and student and benefits to both networks. Kundu et al. tried to extend cross-model distillation to multiple spatially-structured prediction tasks by using two regularization strategies to minimize the domain discrepancy. Yao et al. presented a dense cross-layer mutual distillation mechanism to train the teacher and student collaboratively from scratch. Inspired by the above knowledge distillation techniques, we propose to learn the scene structure guidance for the single DSR task via our designed cross-task distillation. Moreover, another difference to the previous works is that the role of teacher and student is dynamically changed.

3. Method

3.1. Network Architecture

Figure 2 shows the overall architecture of our framework, which mainly consists of three components: depth super-resolution network (DSRNet), depth estimation network (DENet) and the middle cross-task interaction module. Given a collection of paired LR-HR depth maps $\{D_{lr}^{(k)}, D_{hr}^{(k)}\}_{k=1}^{M}$ with the corresponding HR color images $\{I^{(k)}\}_{k=1}^{M}$ as training data, where M is the number of training data, our goal is to learn a model, i.e., DSRNet, that can predict the super-resolved depth maps $\{D_{sr}^{(k)}\}_{k=1}^{M}$ from their corresponding downsampled versions $\{D_{lr}^{(k)}\}_{k=1}^{M}$.

Specifically, the structure of DSRNet is designed based on a network unit from deep back-projection network (DBPN), which can effectively improve the feature representations through iterative projecting HR representations to LR spatial domain and then mapping back into HR domain. The shallow features $F_{sr}^{shallow}$ are first extracted from D_{lr} through a simple convolutional block (including three convolutional layers), and then sent into N stacked DBPN blocks to obtain HR features $\{F_{sr}^{n}\}_{n=1}^{N}$. D_{sr} is finally reconstructed from F_{sr}^{N} through another convolutional block. DENet takes I as input and estimates the depth map D_{de}. The architecture of DENet is similar to DSRNet, but replaces the DBPN blocks with deeper residual blocks to extract informative features $\{F_{de}^{n}\}_{n=1}^{N}$ from color image.

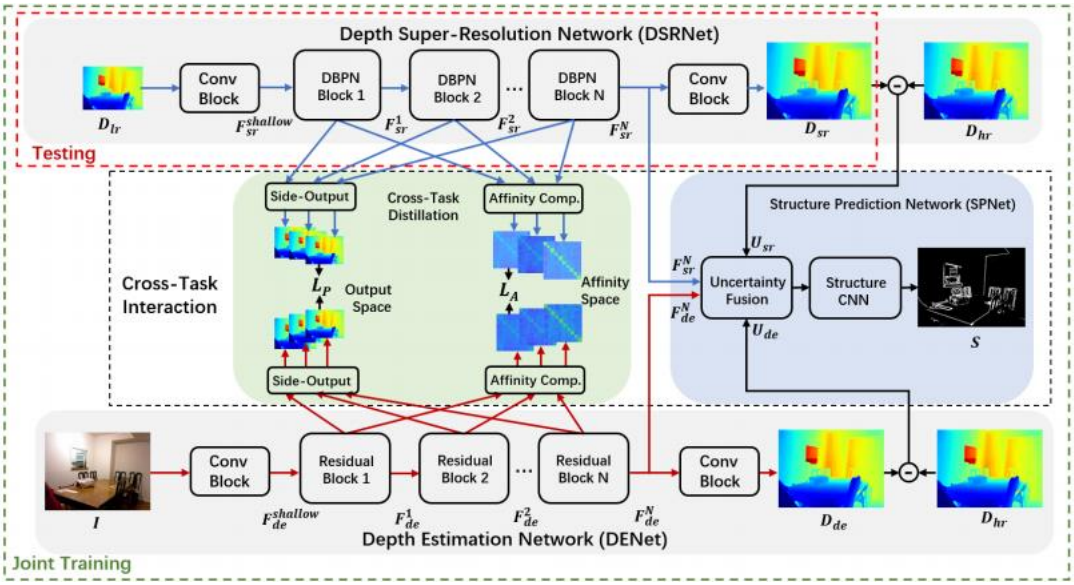

Figure 2 Illustration of our proposed framework, which consists of DSRNet, DENet, and the middle cross-task interaction module. We supervise the outputs of DSRNet and DENet with the same groundtruth depth map D_{hr}. In testing phase, DSRNet is the final choice to predict HR depth map from only LR depth map without the help of color image.

The cross-task interaction module acts as a bridge to connect DSRNet and DENet, and realizes bilateral knowledge transfer between them. It consists of two components, i.e., a cross-task distillation scheme and a structure prediction network (SPNet), where the former focuses on the interaction between multi-scale features extracted from both networks while the latter uses structure maps as supervision to further guide the learning of both networks.

Note that, at the training stage, DSRNet, DENet and the cross-task interaction module are jointly learned by using both color image and depth map as input. In testing phase, DSRNet is the final choice to predict HR depth map from only LR depth map without the help of color image.

3.2. Cross-Task Distillation

Knowledge distillation is generally viewed as a technique of transferring beneficial information from a top-performing model to the other naive one. Different from the commonly used distillation techniques, in which the teacher network is trained beforehand and fixed under the assumption that it always learns a better representation than the student network, our goal is to train SRNet and DENet collaboratively and encourage them to benefit from each other. Inspired by mutual learning methods, we propose a cross-task distillation scheme, in which the roles of teacher and student will exchange between both tasks based on their current performances on depth recovery in the iterative collaborative training. Specially, at the current round of training, we need to determine the teacher in advance according to their performance at the previous round. We compute the average pixel error between each recovered depth map and its ground truth for both networks:

$$e_{dsr} = \frac{1}{HW}\sum_{h}^{H}\sum_{w}^{W}|D_{sr}(h,w) - D_{hr}(h,w)|, \tag{1}$$

$$e_{de} = \frac{1}{HW}\sum_{h}^{H}\sum_{w}^{W}|D_{de}(h,w) - D_{hr}(h,w)|, \tag{2}$$

where {H, W} are the size of the output depth map. If e_{dsr} is smaller than e_{de}, DSRNet has a relatively better performance, and becomes the dominant one to guide the learning of DENet, and vice versa. Next, in order to distill more meaningful knowledge that can accurately describe the essential structural characteristics of depth map, we introduce a multi-space distillation scheme to condense the knowledge from the perspectives of output and affinity spaces, as shown in Figure 2.

Output Space Distillation. To ensure the transfer of local information from pixel-wise depth values in a depth map, we apply the side-output layer (containing two successive convolutions) on $\{F_{sr}^n, F_{de}^n\}_{n=1}^N$ from both DSRNet and DENet to generate the corresponding multi-scale depth outputs $\{D_{sr}^n, D_{de}^n\}_{n=1}^N$

respectively. Thus, the distillation loss of output space is designed to indirectly align the features between DSRNet and DENet:

$$\mathcal{L}_O = \frac{1}{N}\sum_{i=1}^{N}\left\|D_{sr}^{i} - D_{de}^{i}\right\|_1, \tag{3}$$

Affinity Space Distillation. Color image and its associated depth map are different representations of the same scene and have strong structural similarity. Pixels with similar appearances in a color image have more chances of belonging to the same object, and should have close depth values. Inspired by [39, 6, 47] that consider the nonlocal correlations to strengthen correlated features between pixels and benefit the depth map recovery, we also transfer non-local structure knowledge on affinity space, which is implemented by computing pair-wise similarities between pixels.

Assuming the dimension of feature F is $w \times h \times c$, the reshape function $\mathbb{R}$ recasts F as $\mathbb{R}(F)$ with the dimension of $wh \times c$. The affinity matrix A is defined as:

$$A(F) = \sigma\big(\mathbb{R}(F) \otimes \mathbb{R}^T(F)\big), \tag{4}$$

where $\sigma(\cdot)$ is the softmax operation, $\otimes$ is the matrix multiplication and T is the transpose operator. The distillation loss of affinity space is defined as the following,

$$\mathcal{L}_A = \frac{1}{N}\sum_{i=1}^{N}\left\|A\big(F_{sr}^{i}\big) - A\big(F_{de}^{i}\big)\right\|_1, \tag{5}$$

The final distillation loss $\mathcal{L}_{\text{distill}}$ is expressed as:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_O + \gamma\mathcal{L}_A. \tag{6}$$

where γ is an adjustment parameter. Note that, $\mathcal{L}_{\text{distill}}$ should be imposed on the training of the student, but not the teacher, which are determined by the errors comparison between e_{dsr} and e_{de}.

3.3. Structure Prediction

The goal of SPNet is to predict a structure map S from the last feature maps F_{sr}^{n}, F_{de}^{n} generated by DSRNet and DENet respectively. Through the supervision with the ground truth structure map S_{gt}, SPNet can provide extra structure regularization to help both DSRNet and DENet learn more informative structure representations to alleviate the problem of RGB-D structure inconsistency. As shown in Figure 3, SPNet consists of a fusion module and a structure CNN, where the latter is a lightweight network with five stacked ‘Conv+ReLU’ layers and a last ’Conv’ layer.

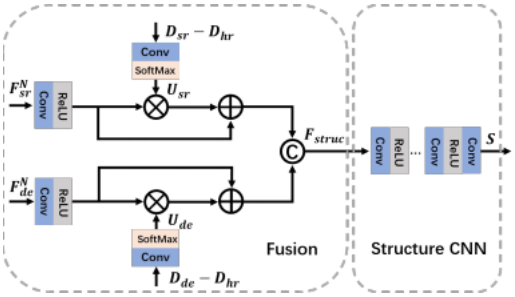

Figure 3. The proposed SPNet. We fuse F_{sr}^N and F_{de}^N by our pro-posed uncertainty-induced attention fusion module.

Usually, the erroneous recovery of DSR and DE tasks occurs in the regions around depth boundaries and fine structures in a depth map, which are subject to higher recovery uncertainty. Therefore, instead of simply concatenating F_{sr}^N and F_{de}^N and sending into the structure CNN, we design an uncertainty-induced attention fusion module to strengthen these informative structure features by attending the recovery uncertainty to the feature map. Thus, we first compute the uncertainty maps U_{sr} and U_{de} of both networks by activating the recovery errors:

$$U_{sr} = \sigma\big(Conv_{1\times1}(D_{sr} - D_{hr})\big) \tag{7}$$

$$U_{de} = \sigma\big(Conv_{1\times1}(D_{de} - D_{hr})\big) \tag{8}$$

where $Conv_{1\times1}$ is the 1×1 convolution to adjust the channels. Then, we use the uncertainty maps to re-weight F_{sr}^N, F_{de}^N and fuse them through an attention module:

$$F_{\text{struc}} = [F_{sr}^N * (1 + U_{sr}), F_{de}^N * (1 + U_{de})] \tag{9}$$

where F_{struc} is the fused features for structure prediction. $[\cdot]$ denotes concatenation operation and $*$ is element-wise multiplication. Note that, through back-propagating the network gradients from SPNet in the backward information flow, the parameters of DSRNet and DENet can be updated.

3.4. Training Algorithm

The training process of our framework can be divided into two steps, as presented in Algorithm 1. First, we separately train DSRNet and DENet with the ground truth D_{hr}. The losses are defined as follows:

$$\mathcal{L}_{DSR} = \|D_{sr} - D_{hr}\|_1, \tag{10}$$

$$\mathcal{L}_{DE} = \lambda\frac{1 - SSIM(D_{de}, D_{hr})}{2} + (1 - \lambda)\|D_{de} - D_{hr}\|_1, \tag{11}$$

where $\mathcal{L}_{DSR}$ is a common pixel-wise L1 loss for the task of DSR. Following [10], $\mathcal{L}_{DE}$ is set as a combination of the reconstruction loss (L1 loss) and structural similarity (SSIM). λ is an adjustment parameter. Then, we introduce the cross-task distillation between both networks with the loss $\mathcal{L}_{\text{distill}}$ in Eq. (6). At the same time, we randomly initialize SPNet, and train it together with DSRNet and DENet. The loss for SPNet is defined as:

$$\mathcal{L}_{\text{struc}} = \left\| \mathbb{G}(F_{\text{struc}}) - S_{gt} \right\|_1 \tag{12}$$

where $\mathbb{G}(\cdot)$ denotes SPNet, F_{struc} is the fused structure feature in Eq.(9) and S_{gt} is the ground truth of the structure. If DSRNet is chosen as the student, the parameters of DENet are fixed at the current epoch, and DSRNet is updated with the following loss:

$$\mathcal{L} = \mathcal{L}_{DSR} + \rho_1 \mathcal{L}_{\text{struc}} + \rho_2 \mathcal{L}_{\text{distill}}, \tag{13}$$

where ρ_1, ρ_2 are the trade-off parameters. Otherwise, DSRNet is fixed, and DENet is updated with the loss $\mathcal{L}$ by replacing $\mathcal{L}_{DSR}$ with $\mathcal{L}_{DE}$.

Algorithm 1 Training Details

Input: Training data D_{lr}, D_{hr}, I, S_{gt}

——————— Step 1 ———————

Randomly initialize DSRNet and DENet

for i = 1; i $\leq$ 100 **do**

  Train DSRNet and DENet with $\mathcal{L}_{DSR}$ and $\mathcal{L}_{DE}$, respectively

——————— Step 2 ———————

Randomly initialize SPNet

for i = 101; i $\leq$ max epoch **do**

  Compute the average error value e_{dsr} and e_{de} according to Eq.(1) and Eq.(2)

  if $e_{dsr} \leq e_{de}$ **then**

    Fix DSRNet and update DENet with

    $\mathcal{L} = \mathcal{L}_{DE} + \rho_1 \mathcal{L}_{struc} + \rho_2 \mathcal{L}_{distill}$

  else

    Fix DENet and update DSRNet with

    $\mathcal{L} = \mathcal{L}_{DSR} + \rho_1 \mathcal{L}_{struc} + \rho_2 \mathcal{L}_{distill}$

Output: D_{sr}

外文翻译

基于跨任务场景结构知识迁移的单张深度图像超分辨率方法

Baoli Sun[1], Xinchen Ye[1,2*], Baopu Li[3], Haojie Li[1,2], Zhihui Wang[1,2], Rui Xu[1,2]

[1]International School of Information Science & Engineering, Dalian University of Technology, China

[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China

[3]Baidu Research, USA

摘要： 现有的色彩引导的深度图像超分辨率重建（DSR）方法需要成对的 RGB-D 数据作为训练样本，其中彩色图像由于其与深度图像的几何相似性而被用作为退化深度图像的恢复提供结构指导。但是，在实际测试中收集配对的数据可能会受到限制或费用昂贵。因此，本文第一次探索了在训练阶段学习跨模态知识，即在训练阶段彩色模态和深度模态的数据对于网络而言都是可见的。但是在测试时，仅存在深度图像一个模态。本文的关键思想是将场景结构指导的知识从彩色模态迁移到到单个深度图像超分辨率重建任务，而无需更改网络结构。具体来说，本文设计了一个用于辅助的深度估计（DE）任务，该任务以彩色图像作为输入来估计深度图像，通过协同训深度图像超分辨率重建任务和深度估计任务以提高深度图像超分辨率重建的性能。在此基础上，提出了跨任务交互模块，实现了跨任务知识的双向传递。首先，本文设计了一种跨任务蒸馏模式，该模式鼓励深度图像超分辨率重建和深度估计网络以师生角色交换的方式相互学习。然后，提出了一种结构预测（SP）任务，该任务提供了额外的结构规范以帮助深度图像超分辨率重建和深度估计网络学习具有更多信息的结构表示以进行深度图像的恢复。大量的实验表明，与其他深度图像超分辨率重建方法相比，本文的方案具有更高的性能。本文代码位于：https://github.com/Sunbaoli/dsr-distillation。

1. 引言

为了更好地理解场景图像，深度信息被补充到彩色图像中，提供了有关场景的关键线索，并在 3D 重建，自动导航，监控等方面取得了广泛的应用。但是，获取室内和室外场景的深度信息需要高昂的成本和巨大的努力，特别是对于高质量和高分辨率（HR）的深度图像。因此，迫切需要一种有效的后处理技术，即深度图像超分辨率重建（DSR），以产生高分辨率的深度图像来缓解此问题。研究人员在深度图像超分辨率重建的方向已经进行了许多努力。通常，由于有限的空间分辨率，在低分辨率（LR）深度图像中，精细的场景结构很容易丢失或严重破坏。 彩色图像及其关联的深度图像分别是同一场景的光度和几何表示，并且具有很强的结构相似性。大多数现有的深度图像超分辨率重建方法从彩色图像中学习结构互补性，以恢复退化的深度图像。

先前的彩色引导的深度图像超分辨率重建方法通过双向融合架构利用 RGB-D 图像对，其中需要一个额外的分支才能从彩色图像中提取结构引导信息。如图 1 (a) 所示，彩色图像和低分辨率深度图像通常由单独的分支处理，并通过融合分支一起滤波以输出高分辨率的结果。但是，由于在网络中间特定层的简单特征融合，在特征提取过程中，来自彩色图像的高频结构信息更容易丢失。因此，如图 1 (b) 所示，一些新颖的方法结合了特征融合的新范式，即多尺度融合，以允许网络学习不同层级的丰富层次特征。这种设计使网络保留了更多的空间细节，以恢复精细和大规模结构。

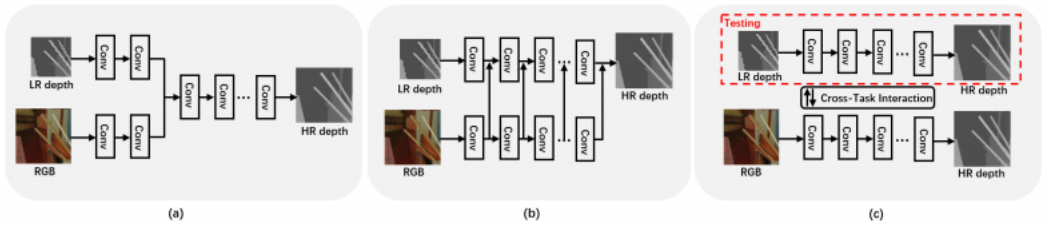

图 1 色彩引导的深度图像超分辨率重建方法示例。(a) 联合滤波，(b) 多尺度特征融合，(c) 本文的跨任务交互机制，可将从彩色图像提的知识蒸馏到深度图像超分辨率重建任务，而无需更改其网络体系结构。

尽管现有的颜色引导的深度图像超分辨率重建方法已显示出显著的进步，但仍然存在一些局限性。首先，这些方法需要成对的 RGB-D 数据作为训练示例，以共同恢复退化的深度图像。但是，在实际测试环境中收集配对数据可能会受到限制或费用昂贵。例如彩色图像和深度图像是分别由单独的具有不同分辨率和视图的深度传感器和彩色传感器捕获的，因此需要在它们之间进行精确的标定和校正才能获得配对的图像对。实际上，大多数实际应用仍然仅有一个低分辨率的深度图像，这引发了上述问题。其次，考虑到内存消耗和计算负担，对高分辨率彩色图像的处理也阻碍了实际应用。此外，尽管可以将彩色特征用作深度图像超分辨率重建中退化图像恢复的结构指导，但彩色图像的结构并不总是与深度图像的结构一致（结构不一致），这会导致明显的伪影，例如纹理复制和深度渗色。因此，如何利用彩色信息来帮助恢复深度图像并同时满足实际的测试环境，仍有待研究。

基于以上分析，本文摆脱了一般范式的束缚，针对深度图像超分辨率重建的任务介绍了一种新颖的场景结构指导学习方法，如图 1 (c) 所示。本文第一次探索了在训练阶段学习跨模态知识，即在训练阶段可以同时使用彩色模态和深度模态的图像，但是在目标数据集上进行测试时，仅存在深度模态的图像。本文的关键思想是在不更改网络结构的情况下，将场景结构指导的知识从彩色模态前一到单独的深度图像超分辨率重建任务。

具体来说，如图 2 所示，受多任务学习成功经验的启发，本文构造了一个辅助的深度估计（DE）任务，该任务将彩色图像作为输入来估计深度图像。在此基础上，本文提出了一个跨任务交互模块，以实现深度图像超分辨率重建任务和深度估计任务之间的双边知识迁移。与常用的蒸馏技术不同，本文首先设计一种跨任务蒸馏，该蒸馏鼓励深度图像超分辨率重建网络（DSRNet）和深度估计网络（DENet）相互学习，即老师和学生的角色将根据两个任务在迭代式协作训练中他们目前在深度图像恢复方面的表现动态切换。本文还引入了多空间蒸馏方案，从输出和亲和空间的角度进行知识蒸馏，从而可以更好地描述深度图像的基本结构特征。此外，为了解决 RGB-D 结构不一致的问题，本文构

造了一个结构预测（SP）任务，该任务提供了额外的结构规范化以帮助 DSRNet 和 DENet 都学习更多有关深度图像恢复信息的结构表示。本文提出了一个不确定度引导的注意力融合模块，为结构预测网络（SPNet）提供合理的输入，然后使用从 DSRNet 和 DENet 获得的不确定性图来对两个网络的特征进行加权，以用于加强有效的结构知识。大量的实验表明，在准确性和运行时间方面，本文的深度图像超分辨率重建方法甚至优于基准数据集上的颜色引导的深度图像超分辨率重建方法。主要贡献总结如下：

- 据作者所知，本文是第一个提出在训练阶段以多种模态作为深度图像超分辨率重建的输入，但在测试时仅以低分辨率深度图像作为输入进行。
- 提出了跨任务的蒸馏方案，以鼓励 DSRNet 和 DENet 在协作训练模式下互相学习。
- 引入了结构预测网络，以提供结构信息的规范化，从而帮助 DSRNet 解决两种模态结构不一致的问题。

2. 相关工作

深度图像超分辨率重建. 与单一的深度图像超分辨率重建方法相比，色彩引导的深度图像超分辨率重建方法已被广泛提出来通过彩色图像的引导改善深度图像的质量。 Li 等提出了一种联合滤波方法，该方法利用彩色图像作为指导来增强深度图像的空间分辨率。 Hui 等采用多尺度融合策略，将不同层次的颜色特征融合在一起，以解决深度图像超分辨率重建中的歧义。 Wen 等提出了一种数据驱动的滤波器，以彩色图像为指导推断初始的高分辨率深度图像，然后提出了从粗到精的网络来逐步恢复深度图像。Guo 等提出了一种分层特征驱动的方法，该方法构造了用于多级残差学习的输入金字塔和引导金字塔。 Wang 等提出利用从彩色图像中学到的边缘图对深度图进行上采样。

单目深度估计. 由于 CNN 在特征提取中的强大功能，有监督的单目深度估计方法在不断地提高深度估计的性能。Laina 等提出了一种全卷积神经网络来对彩色图像和深度图像之间的映射进行建模。在[25]中，提出了一种双分支的 CNN 来同时预测深度图像和深度梯度，以进行准确的深度估计。 Wang 等提出了一种基于语义分治策略的深度估计网络，将场景分解为语义片段，然后预测每个片段的深度。与之不同的是，无监督方法在训练过程中使用视频或立体图像对，而无需深度图像的真值。 Wong 等通过双分支解码器来学习一种用于估计深度图像的鲁棒表示。

知识蒸馏. 知识蒸馏是将知识从大体量模型转移到轻量模型，以提高后者的性能。它已被广泛应用于许多领域中，包括动作识别，风格转换，深度估计和场景理解。例如，图像分类任务将来自教师网络的类别概率作为软目标来训练学生网络或通过中间层传递知识。最近，深度互学习提出了两路蒸馏方法，在老师和学生之间传递知识并使两个网络都受益。Kundu 等通过使用两种正则化策略将跨模型蒸馏扩展到多个空间结构的预测任务，以最小化域差异。Yao 等提出了一种密集的跨层相互蒸馏机制，以从头开始协作训练教师与学生。受以上知识蒸馏技术的启发，本文提出跨任务蒸

馏学习深度图像超分辨率重建任务的场景结构指导。而且，与以前的工作的另一个区别是教师和学生的角色是动态变化的。

3. 方法

3.1. 网络架构

图 2 显示了本文网络的总体架构，主要由三个部分组成：深度超分辨率网络（DSRNet），深度估计网络（DENet）和中间的跨任务交互模块。给定成对的 LR-HR 深度图像 $\{D_{lr}^{(k)}, D_{hr}^{(k)}\}_{k=1}^{M}$ 以及相应的高分辨率彩色图像 $\{I^{(k)}\}_{k=1}^{M}$ 作为训练数据，其中 M 是训练数据的数量，其目标是学习模型，即 DSRNet，可以根据对应的降采样版本 $\{D_{lr}^{(k)}\}_{k=1}^{M}$ 恢复超分辨深度图像 $\{D_{sr}^{(k)}\}_{k=1}^{M}$。

具体而言，DSRNet 的结构是基于来自深度反投影网络（DBPN）的网络单元设计的，它可以通过将高分辨率表示迭代地投影到低分辨率空间域然后再映射回高分辨率域来有效地改善特征表示。首先通过简单的卷积块（包括三个卷积层）从 D_{lr} 中提取浅层特征 $F_{sr}^{shallow}$，然后将其发送到 N 个堆叠的 DBPN 块中以获得高分辨率特征 $\{F_{sr}^{n}\}_{n=1}^{N}$。最后通过另一个卷积块从 F_{sr}^{N} 重建 D_{sr}。DENet 将 I 作为输入，并估计深度图像 D_{de}。 DENet 的结构类似于 DSRNet，但是用更深的残差块替换了 DBPN 块，以从彩色图像中提取更具信息的特征 $\{F_{de}^{n}\}_{n=1}^{N}$。

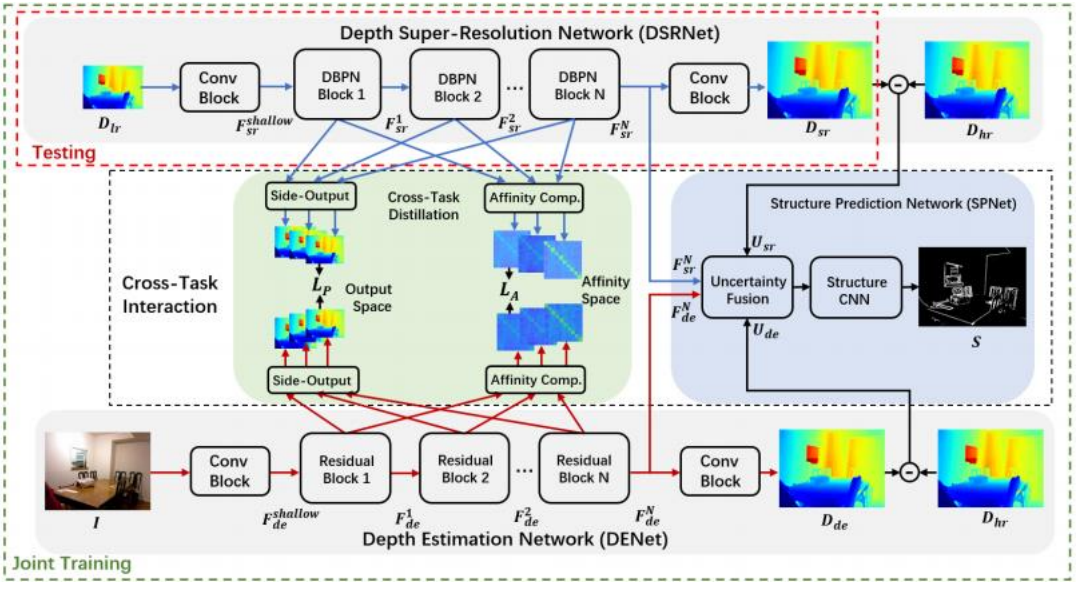

图 2 本文提出的网络架构，该框架由 DSRNet，DENet 和中间的跨任务交互模块组成。本文使用相同的深度图像真值 D_{hr} 监督 DSRNet 和 DENet 的输出。在测试阶段，DSRNet 是仅从低分辨率深度图像而不借助彩色图像预测的高分辨率深度图像的最终选择。

跨任务交互模块充当连接 DSRNet 和 DENet 的桥梁，并实现它们之间的双边知识迁移。它由跨任务蒸馏模式和结构预测网络（SPNet）组成，其中前者着重于两个网络提取的多尺度特征之间的交互，而后者则使用结构图作为监督信息来进一步指导两个网络的学习。

需要注意的是，在训练阶段，通过使用彩色图像和深度图像作为输入来共同训练 DSRNet、DENet 和跨任务交互模块。在测试阶段，DSRNet 不借助彩色图像而仅从低分辨率深度图像恢复高分辨率深度图像。

3.2. 跨任务知识蒸馏

知识蒸馏通常被视为一种将有益的信息从性能高的模型迁移到另一个简略的模型的技术。常用的知识蒸馏技术事先对教师网络进行训练并在其始终比学生网络学得更好的假设下固定其参数，与之不同，而本文的目标是协作训练 SRNet 和 DENet 并鼓励他们彼此受益。受相互学习方法的启发，本文提出了一个跨任务的知识蒸馏策略，其中

在迭代式协作训练中，老师和学生的身份将根据当前在深度图像恢复上的表现在两个任务之间进行交换。具体而言，在每一轮训练中，需要根据上一轮的表现提前确定教师。分别计算两个网络恢复的深度图像与深度图像真值之间的平均像素误差：

$$e_{dsr} = \frac{1}{HW}\sum_{h}^{H}\sum_{w}^{W}|D_{sr}(h,w) - D_{hr}(h,w)|, \tag{1}$$

$$e_{de} = \frac{1}{HW}\sum_{h}^{H}\sum_{w}^{W}|D_{de}(h,w) - D_{hr}(h,w)|, \tag{2}$$

其中 {H, W} 是输出深度图像的大小。如果 e_{dsr} 小于 e_{de}，则 DSRNet 具有相对较好的性能，并成为指导 DENet 学习的教师，反之亦然。接下来，为了提取更有意义的知识以准确描述深度图像的基本结构特征，本文引入了一种多空间蒸馏策略，从输出和亲和空间的角度来浓缩知识，如图 2 所示。

输出空间蒸馏. 为了确保深度图像中局部信息的像素级深度值迁移，在来自 DSRNet 和 DENet 的 $\{F_{sr}^{n}, F_{de}^{n}\}_{n=1}^{N}$ 上应用侧面输出层（包含两个连续的卷积），分别生成相应的多尺度深度图像输出 $\{D_{sr}^{n}, D_{de}^{n}\}_{n=1}^{N}$。因此，输出空间的蒸馏损失旨在间接调整 DSRNet 和 DENet 之间的特征：

$$\mathcal{L}_{O} = \frac{1}{N}\sum_{i=1}^{N}\left\|D_{sr}^{i} - D_{de}^{i}\right\|_{1}, \tag{3}$$

亲和空间蒸馏. 彩色图像及其关联的深度图像是同一场景的不同表示，并且具有很强的结构相似性。在彩色图像中具有相似外观的像素更有可能属于同一目标，并且应该具有接近的深度值。受工作 [39，

6，47] 的启发，这些研究考虑了非局部相关性以增强像素之间的相关特征并有利于深度图像的恢复，因此本文还在亲和空间上传递非局部的结构知识，其是通过计算成对像素之间的相似性来实现的。

假设维度为 $w \times h \times c$ 的特征，则整形函数 $\mathbb{R}$ 将 F 重塑为维度为 $wh \times c$ 的特征 $\mathbb{R}(F)$。亲和矩阵 A 定义为：

$$A(F) = \sigma\big(\mathbb{R}(F) \otimes \mathbb{R}^T(F)\big), \tag{4}$$

其中 $\sigma(\cdot)$ 是 softmax 运算，$\otimes$ 是矩阵乘法，T 是转置运算符。亲和空间的蒸馏损失定义如下：

$$\mathcal{L}_A = \frac{1}{N}\sum_{i=1}^{N}\left\|A\left(F_{sr}^i\right) - A\left(F_{de}^i\right)\right\|_1, \tag{5}$$

最终蒸馏损失 $\mathcal{L}_{\text{distill}}$ 表示为：

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_O + \gamma\mathcal{L}_A. \tag{6}$$

其中 γ 是调整参数。需要注意的是，$\mathcal{L}_{\text{distill}}$ 应该施加给学生网络，而不是教师网络，这是由 e_{dsr} 和 e_{de} 之间的误差比较确定的。

3.3. 结构预测

SPNet 的目标是根据分别由 DSRNet 和 DENet 生成的最后一个特征图 F_{sr}^n, F_{de}^n 预测结构图 S。通过使用结构图真值 S_{gt} 进行监督，SPNet 可以提供额外的结构规范化，以帮助 DSRNet 和 DENet 都学习具有更多信息的结构表示，从而缓解 RGB-D 结构不一致的问题。如图 3 所示，SPNet 由一个融合模块和一个结构 CNN 组成，其中 CNN 是一个轻量级网络，由五个堆叠的“Conv＋ReLU”层和最后的一个“Conv”层组成。

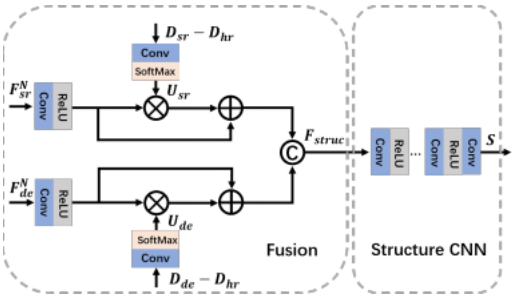

图 3 本文提出的 SPNet。通过提出的不确定度引导的注意融合模块将 F_{sr}^N 和 F_{de}^N 融合。

通常，DSR 和 DE 任务的恢复的误差往往出现在深度边界和精细结构附近区域，这些区域具有较高的恢复不确定性。因此，不是简单地将 F_{sr}^N 和 F_{de}^N 级联并馈入到 CNN 结构中，而是设计了一个不确定度引导的注意融合模块，通过将恢复不确定度纳入特征图来增强这些具有信息性的结构特征。因此，本文首先通过激活恢复误差来计算两个网络的不确定度图 U_{sr} 和 U_{de}：

$$U_{sr} = \sigma\big(Conv_{1\times1}(D_{sr} - D_{hr})\big) \tag{7}$$

$$U_{de} = \sigma\big(Conv_{1\times1}(D_{de} - D_{hr})\big) \tag{8}$$

其中 $Conv_{1\times1}$ 是用于调整通道的 1×1 卷积。然后，使用不确定度图对 F_{sr}^{N}, F_{de}^{N} 进行重新加权并通过注意力模块将它们融合：

$$F_{\text{struc}} = [F_{sr}^{N} * (1 + U_{sr}), F_{de}^{N} * (1 + U_{de})] \tag{9}$$

其中 F_{struc} 是用于结构预测的融合特征。 [·]表示级联运算，而 * 是元素级相乘。需要注意的是，通过向后信息流在 SPNet 中梯度的反向传播，可以更新 DSRNet 和 DENet 的参数。

3.4. 训练策略

本文网络的训练过程可以分为两个步骤，如算法 1 所示。首先，分别使用深度图像的真值 D_{hr} 训练 DSRNet 和 DENet。损失定义如下：

$$\mathcal{L}_{DSR} = \|D_{sr} - D_{hr}\|_1, \tag{10}$$

$$\mathcal{L}_{DE} = \lambda\frac{1 - SSIM(D_{de}, D_{hr})}{2} + (1 - \lambda)\|D_{de} - D_{hr}\|_1, \tag{11}$$

其中 $\mathcal{L}_{DSR}$ 是深度图像超分辨率重建任务的常见像素级 L1 损失。根据工作[10]，将 $\mathcal{L}_{DE}$ 设置为重建损失（L1 损失）和结构相似性（SSIM）损失的组合。λ 是调整参数。然后，引入了在等式（6）中介绍的两个网络之间的交叉任务蒸馏损失 $\mathcal{L}_{\text{distill}}$。同时，本文随机初始化 SPNet，并将其与 DSRNet 和 DENet 一起训练。SPNet 的损失定义为：

$$\mathcal{L}_{\text{struc}} = \left\|\mathbb{G}(F_{\text{struc}}) - S_{gt}\right\|_1 \tag{12}$$

其中 $\mathbb{G}(\cdot)$ 表示 SPNet，F_{struc} 是等式（9）中的融合结构特征，而 S_{gt} 是该结构的真值。如果选择 DSRNet 作为学生，则 DENet 的参数将被固定为当前轮次的值，然后对 DSRNet 进行更新，有以下损失：

$$\mathcal{L} = \mathcal{L}_{DSR} + \rho_1\mathcal{L}_{\text{struc}} + \rho_2\mathcal{L}_{\text{distill}}, \tag{13}$$

其中 ρ_1, ρ_2 是平衡参数。否则，DSRNet 的参数是固定的，并且通过将 $\mathcal{L}_{DSR}$ 替换为 $\mathcal{L}_{DE}$，以损失 $\mathcal{L}$ 更新 DENet。

Algorithm 1 Training Details

Input: Training data D_{lr}, D_{hr}, I, S_{gt}

——————— Step 1 ———————

Randomly initialize DSRNet and DENet

for i = 1; i ≤ 100 **do**

  Train DSRNet and DENet with $\mathcal{L}_{DSR}$ and $\mathcal{L}_{DE}$, respectively

——————— Step 2 ———————

Randomly initialize SPNet

for i = 101; i ≤ max epoch **do**

  Compute the average error value e_{dsr} and e_{de} according to Eq.(1) and Eq.(2)

  if $e_{dsr} \leq e_{de}$ **then**

    Fix DSRNet and update DENet with

    $\mathcal{L} = \mathcal{L}_{DE} + \rho_1\mathcal{L}_{struc} + \rho_2\mathcal{L}_{distill}$

  else

    Fix DENet and update DSRNet with

    $\mathcal{L} = \mathcal{L}_{DSR} + \rho_1\mathcal{L}_{struc} + \rho_2\mathcal{L}_{distill}$

Output: D_{sr}