DOI:10.19651/j.cnki.emt.2107618

# 改进卷积空间传播网络的单目图像深度估计*

刘安旭 黎向锋 刘晋川 王建明 赵 康 左敦稳
(南京航空航天大学 机电学院 南京 210016)

**摘 要:** 单目图像深度估计是计算机视觉领域中的一个基本问题,卷积空间传播网络(CSPN)是现阶段最先进的单目图像深度估计方法之一。针对 CSPN 在预测密集深度图时存在部分物体结构变形和物体间边缘模糊不清的边界混合问题,分别从网络结构与损失函数两部分进行了改进。对输入稀疏深度图进行了 3 次不同尺寸下采样,并将其加入到 U-Net 模块相应的编码过程和跳跃连接部分,以使其能够更精确地捕捉不同尺度的物体结构。并使用深度误差对数、深度信息梯度及表面法线这 3 种损失函数加权组合形成的改进损失函数来替换原始损失函数。在 NYU-Depth-V2 数据集上的实验结果表明,改进卷积空间传播网络(ICSPN)与 CSPN 相比,其均方根误差 RMSE 降低了 17.23%,平均相对误差 REL 降低了 28.07%。ICSPN 充分利用了输入稀疏深度图,减小了预测密集深度图中物体结构的变形,同时采用带有梯度损失的损失函数对训练过程进行监督,降低了物体边缘位置误差,减少了边界混合问题的产生。

**关键词:** 深度估计;单目视觉;深度学习;卷积空间传播网络;梯度损失
**中图分类号:** TP391.4 **文献标识码:** A **国家标准学科分类代码:** 520.6030

# Monocular image depth estimation of improved convolutional spatial propagation network

Liu Anxu Li Xiangfeng Liu Jinchuan Wang Jianming Zhao Kang Zuo Dunwen
(College of Mechanical and Electrical Engineering, Nanjing University of Aeronautic and Astronautics, Nanjing 210016, China)

**Abstract:** Monocular image depth estimation is a basic problem in the field of computer vision, convolutional spatial propagation network (CSPN) is one of the most advanced monocular image depth estimation methods. Aiming at the deformation problem of some objects and the boundary mixing problem caused by the blurring of the edges between objects in the dense depth map predicted by the network, we have improved CSPN from the network structure and loss function respectively. The input sparse depth map is downsampled three times with different sizes and added to the corresponding coding process and skip connection part of the U-Net module, so that it can more accurately capture the structure of objects with different scales. The original loss function is replaced by the improved loss function formed by the weighted combination of depth error logarithm, depth information gradient and surface normal. The experimental results on NYU-Depth-V2 data set show that compared with CSPN, the root mean square error RMSE and average relative error REL of ICSPN are reduced by 17.23% and 28.07% respectively. The ICSPN makes full use of the input sparse depth map to reduce the deformation of the object structure in the predicted dense depth map. At the same time, the loss function with gradient loss is used to monitor the training process, which reduces the edge position error of the object and the problem of boundary mixing.

**Keywords:** depth estimation;monocular vision;deep learning;convolutional spatial propagation network;gradient loss

## 0 引 言

深度估计是用来预测图像中每个像素点到相机的距离,在自动驾驶、机器人导航和增强现实等领域有着广泛的应用。深度估计旨在获取 RGB 图像对应的深度图像,其深度图像中每个像素点的灰度值可用于表征场景中某一点距离相机的远近。通常包括基于多视点图像的深度估计、基于双目图像的深度估计和基于单目图像的深度估计[1]。其中基于多视点图像的深度估计是对同一场景通过多台相机进行多角度拍摄,并利用多视点图像之间的冗余信息进行

收稿日期:2021-08-17
* 基金项目:国家自然科学基金联合基金项目(U20A20293)资助

深度信息的计算。这类技术通常能够获得比较准确的深度信息,但是由于需要对多台相机进行特定位置的排列,在大多数实际应用中很少采用。基于双目图像的深度估计是运用与人双眼相对位置相同的两台相机对同一场景成像,并通过立体匹配技术实现深度信息的计算。基于单目图像的深度估计是以单台相机作为感知信息的输入,通过一系列算法流程和图像处理,预测当前 RGB 图像对应的深度图。与前两种深度估计相比,单目图像深度估计对硬件要求低,具有轻便、廉价、易于移植拓展与落地的特点,因此在实际生产环境中有更广阔的应用前景。

传统单目图像深度估计的方法主要是基于场景中的一些视觉线索来估计深度,其方法主要包括:从明暗度变化的规律恢复形状(shape from shading,SFS[2])、从运动中恢复形状(shape from motion,SFM[3])、从对焦获取深度(depth from focus,DFF)及从离焦获取深度(depth from defocus,DFD[4])等。利用这些深度线索可以推断图像中的场景结构以及深度相关信息,并进而估计出场景深度。但这些方法都是建立在特殊环境条件下或是需要较为明显的辅助设备,且最终结果受环境因素影响很大,使用条件苛刻难以推广。因此,需要研究适用范围更广、性能更加鲁棒、环境条件要求低以及使用更加便捷的单目图像深度估计方法。

近年来,随着深度学习的普及与发展,利用神经网络模型得到图像深度逐渐成为主流研究方向[5]。在此情境下,学者们提出了各种基于深度学习的单目图像深度估计算法。根据网络模型输入对象的不同主要分为两种,一种是输入单幅 RGB 图像直接预测输出深度图,主要应用于早期的深度学习方法中,如 Eigen 等[6]首次提出使用卷积神经网络(convolutional neural network,CNN)估计单张彩色图像对应的深度信息,整体网络框架由 1 个全局粗糙网络和 1 个局部精细网络堆叠组成。Liu 等[7]提出将连续形式的条件随机场(continuous conditional random field, CCRF)与卷积神经网络相结合,构建统一的神经网络以完成单目图像的深度预测。Fu 等[8]提出一种间距增大的离散化策略(statistical independence discretization,SID)将深度离散化,用一个序数回归损失函数训练网络模型。同时,采用语义分割领域 DeepLab 系列提出的空洞空间金字塔池化(atrous spatial pyramid pooling,ASPP)结构并行提取多尺度信息,避免不必要的空间池化操作。

虽然单目深度估计可以只利用 RGB 图像预测相应的深度图,但是因为缺乏相应的先验深度信息(如由激光雷达传感器获得的可靠精确的深度信息),无法估计出高质量的密集深度图。因此,另一种则是输入为单幅 RGB 图像和相应的稀疏深度图,输出为密集深度图,这种方法由于获得了稀疏深度图的引导,预测的密集深度图更为准确,本文所研究的输入对象即为单幅 RGB 图像和相应的稀疏深度图。如 Liu 等[9]提出了一种学习局部相似性的空间传播网络(spatial propagation network,SPN)。这种 SPN 可以从大规模数据中学习特定任务的相似性矩阵,并结合稀疏深度图来引导密集深度图的生成,实验效果取得了显著的提升。但是其空间传播模块采用的是单个的四方向三向连接,不适合同时考虑所有的局部邻域。Cheng 等[10]克服了这一局限,提出了一种卷积空间传播网络(convolutional spatial propagation network,CSPN),用于预测局部邻域的相似性矩阵,并利用其局部上下文同时更新所有像素以提高效率,取得了较好的效果。但是该网络预测的密集深度图中会产生部分物体结构变形问题和物体间边缘模糊不清导致的边界混合问题。

基于上述问题,本文对 CSPN 进行改进,在其 U-Net 模块对原始稀疏深度图进行下采样使之引导 U-Net 模块的编码及解码过程,另一部分使用改进的损失函数来对训练过程进行监督,以减少物体结构变形和边界混合现象的出现,进一步提高其预测精度。

## 1 卷积空间传播网络

### 1.1 空间传播网络(SPN)

Liu 等[9]提出通过深度卷积神经网络(deep convolutional neural networks,DCNN)学习相似性矩阵和空间传播模块细化输出,取得了较好的预测效果。SPN 可以将二维图片(如原始 RGB 图和稀疏深度图)转换为具有所需属性的图像(如密集深度图)。SPN 建立了一个三向连接的线性传播模型,如图 1(b)所示,深度图中的像素值按照 4 个方向先后顺序依次扫描更新,每次扫描时当前像素点根据此方向指示的相邻 3 个像素点进行计算得到。相比图 1(a)单向连接产生的全局稀疏连接的成对关系,三向连接能够形成全局的密集连接的成对关系。

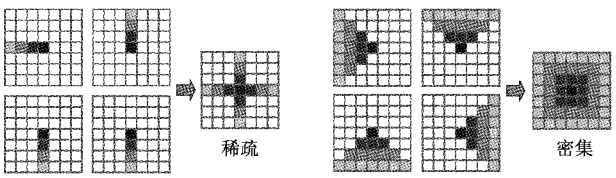



图 1 单向连接与三向连接的不同传播范围

SPN 使用三向连接来细化输出,图 2 为 SPN 的网络结构,网络由黑色虚线分隔,包含传播模块(上部)和制导网络(下部),其中制导网络输出 4 个方向传播的相似性矩阵,相似性矩阵是一个通用矩阵,用于确定空间中两点的接近程度或相似程度。在计算机视觉任务中,它是一个加权图,将每个像素视为一个节点,并通过边连接每对像素[9]。稀疏深度图和相似性矩阵在传播模块作用下输出所需的密集深度图。

### 1.2 卷积空间传播网络(CSPN)

虽然 SPN 相较于之前的方法,在预测精度和准确度上有了较大地提升,但是它的传播是以扫描行或扫描列的方式进行的,这在本质上是串行的。例如,当从左向右传播

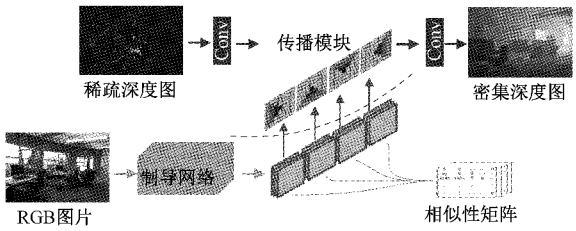



图 2　SPN 网络结构

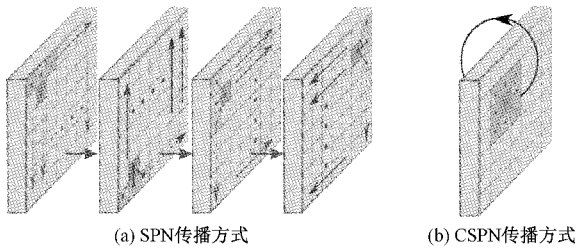



图 3　SPN 与 CSPN 传播方式对比

时，最右列的像素必须等待最左列的信息来更新其值。针对此问题，Cheng 等[10]提出了一种卷积空间传播网络 CSPN，不同于 SPN 中三向连接按 4 个方向先后顺序扫描整个图像，CSPN 使用 $k\times k$ 卷积核同时进行四方向传播计算，（如图 3(b)中的 $k\times k$ 为 $3\times 3$，其中心点的像素值是由 $3\times 3$ 卷积核计算得到），而在执行重复处理时可以获得范围更大的全局密集连接关系。CSPN 的并行更新方案在速度和质量方面都比 SPN 串行更新方案有了显著的提高。

不同于 Liu 等[9]建立的学习相似性矩阵的深层网络，CSPN 从给定网络分支出一个额外的相似性矩阵预测网络，它与密集深度预测网络共享相同的特征提取网络，用于预测相似性矩阵，这有助于为密集深度估计和相似性矩阵预测的联合学习节省内存和时间成本，如图 4 所示。相似性矩阵的学习依赖于输入图像的细粒度空间细节，然而，在编码器特征提取过程中，随着下采样操作的进行，其空间信息会减弱或丢失。因此，CSPN 网络添加了类似于 U-Net 的镜像连接，将编码器提取的特征跳跃连接到上采样操作部分。其中，UpProj 模块用于不含跳跃连接部分的上采样，UpProj_Cat 用于包含跳跃连接部分的上采样，此部分输出初步生成的深度图和相似性矩阵。最后，输入稀疏深度图、初步生成的深度图和相似性矩阵在 CSPN 模块的作用下输出最终预测的密集深度图。

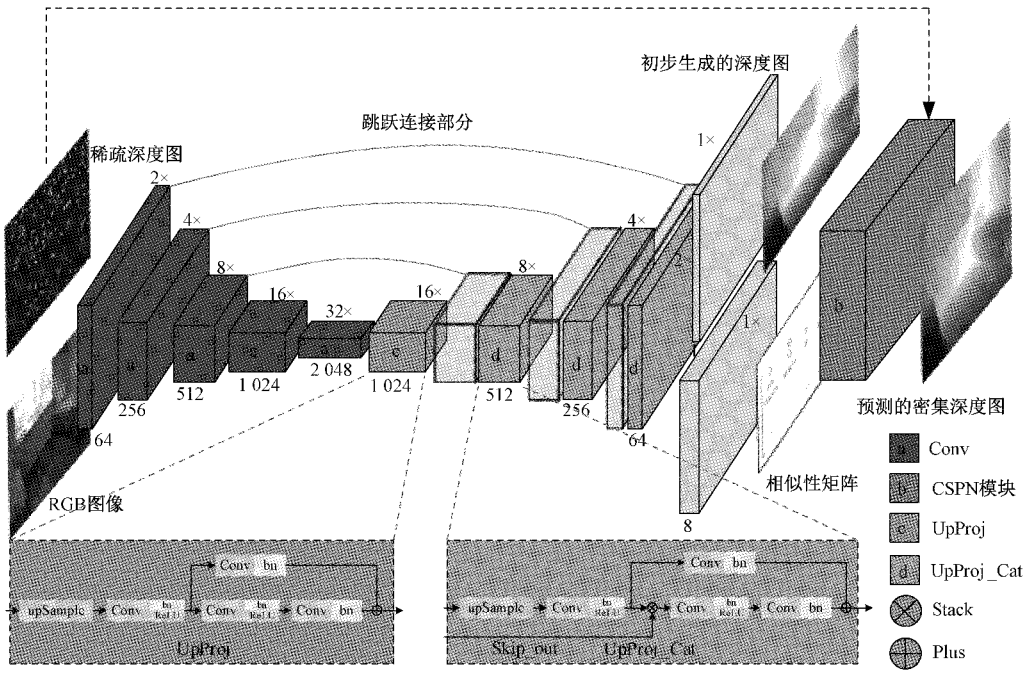



图 4　CSPN 网络结构

### 1.3　稀疏深度图的获取

由于 NYU-V2 数据集中只包含 RGB 图像及对应的密集深度图，因此需要对密集深度图 $\boldsymbol{D}$ 进行采样以获得其稀疏深度图 $s\boldsymbol{D}$，$s\boldsymbol{D}$ 用来模拟一般激光雷达传感器获得的稀疏深度信息。在网络训练过程中，输入稀疏深度图 $s\boldsymbol{D}$ 是从真实密集深度图 $\boldsymbol{D}$ 中随机采样，深度图中的非采样点像素值均为 0。如，对于任意采样像素数量 $m$（在训练过程是固定的）的稀疏深度样本，计算其伯努利概率 $p=m/n$，其中 $n$ 是 $\boldsymbol{D}$ 中有效深度像素的总数。对于任意位置的像素 $(i,j)$，则：

$$s\boldsymbol{D}(i,j)=\begin{cases}\boldsymbol{D}(i,j), & \text{概率为 } p\\ 0, & \text{概率为 } 1-p\end{cases} \tag{1}$$

使用这种采样方法，每个稀疏深度样本的非零深度像素的实际数量在期望值 $m$ 附近变化[11]。这种采样方法可

以增加网络对不同采样点数量的稀疏深度输入的鲁棒性，并增加训练集的多样性。

## 2　本文的网络结构和损失函数

### 2.1　网络结构

深度估计技术可以广泛应用于机器人和自动驾驶领域，在这些领域中深度信息往往是通过激光雷达获取，这通常会产生稀疏但精确的深度信息。但这些高度可靠的深度信息在 CSPN 中并未得到充分利用，CSPN 中稀疏深度图只作为 U-Net 模块和 CSPN 模块的输入进行训练。因此，本文将原始大小的稀疏深度图分别进行不同尺寸大小的下采样，并加入到 U-Net 模块的编码过程和跳跃连接部分，以引导初步生成的深度图与相似性矩阵的生成，如图 5 中 U-Net 模块的虚线和 Cat_Conv 模块所示。

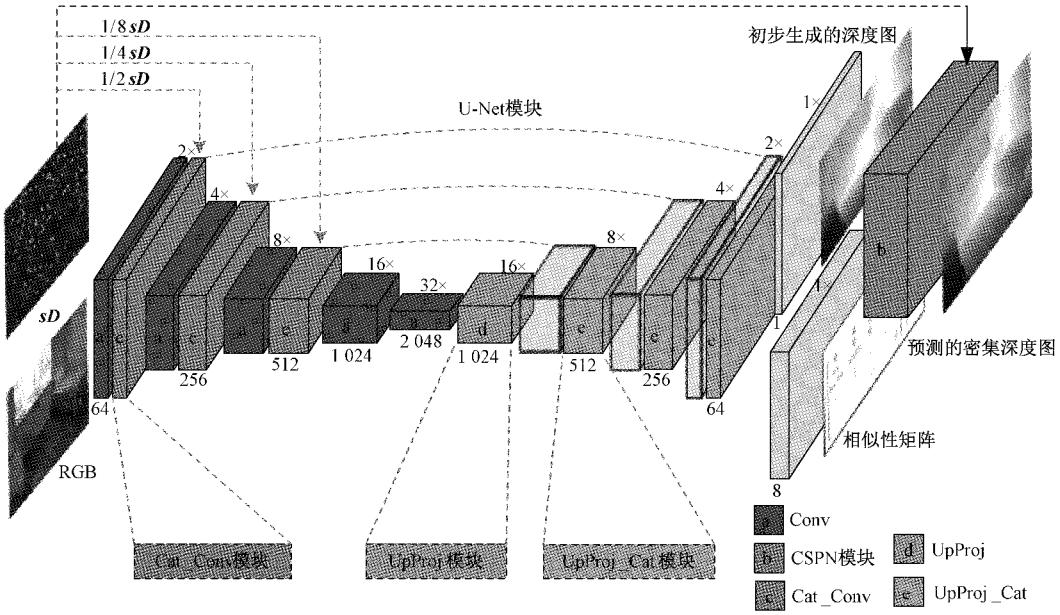



图 5　本文的 ICSPN 网络结构

在对稀疏深度图进行下采样时，要保留尽可能多的信息，因此采用标准平均池化对稀疏深度图进行下采样[12]。下采样稀疏深度图 $\boldsymbol{sD}^k$ 上位置$(x,y)$处的数据是原始稀疏深度图 $\boldsymbol{sD}$ 上像素$(2^k x,2^k y)$的有效邻域的平均值，其中 $2^k$ 为下采样因子。这个下采样操作是原始稀疏深度图 $\boldsymbol{sD}$ 的平均池化结果除以掩膜 $\boldsymbol{C}$ 的平均池化结果(处理流程如图 6 所示)，其中 $\boldsymbol{C}(x,y)=1$ 表示 $\boldsymbol{sD}$ 中的像素$(x,y)$不为 0，否则 $\boldsymbol{C}(x,y)=0$。下采样操作 $\varphi^k_{x,y}(\boldsymbol{sD},\boldsymbol{C})$ 的过程如式(2)所示。

$$\varphi^k_{x,y}(\boldsymbol{sD},\boldsymbol{C})=\frac{\sum_{i,j=0}^{2^k-1}\boldsymbol{sD}_{2^k x+i,2^k y+j}}{\sum_{i,j=0}^{2^k-1}\boldsymbol{C}_{2^k x+i,2^k y+j}+\varepsilon}=\frac{\sum_{i,j=0}^{2^k-1}\boldsymbol{sD}_{2^k x+i,2^k y+j}/2^{2k}}{\sum_{i,j=0}^{2^k-1}\boldsymbol{C}_{2^k x+i,2^k y+j}/2^{2k}+\varepsilon}=\frac{\phi^k_{x,y}(\boldsymbol{sD})}{\phi^k_{x,y}(\boldsymbol{C})+\varepsilon} \tag{2}$$

其中，$\varphi^k_{x,y}(\cdot)$为平均池化操作，$\varepsilon$ 是避免分母为 0 设置的很小数。

首先，将输入的稀疏深度图 $\boldsymbol{sD}$ 下采样至其原始尺寸的 1/2(记为 $1/2\boldsymbol{sD}$)，并与原始 RGB 图及其稀疏深度图拼接后的下采样特征图(图 6 中的 $\boldsymbol{X}$)进行再次拼接(记为 Cat)及融合(图 6 中的 Conv)，以引导编码器更高层次的特征提取；同时通过跳跃连接将卷积融合后的结果传递至解码器，为解码器的上采样部分提供更多细节信息。之后，用相同方法将 $\boldsymbol{sD}$ 下采样至 $1/4\boldsymbol{sD}$ 和 $1/8\boldsymbol{sD}$，并加入到 U-Net 模块，以便使 U-Net 模块能够更精确地捕捉到场景中较大的结构，本文 U-Net 模块整体结构如图 7 所示。

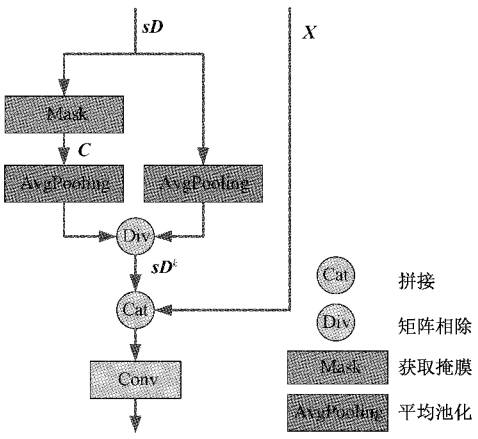



图 6　$\boldsymbol{sD}$ 下采样与特征图 $X$ 融合的网络结构

### 2.2　损失函数

以前的大多数研究都采用深度估计值与其真实值之

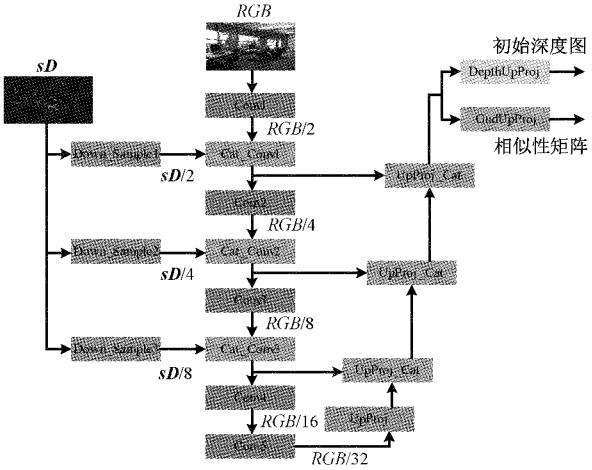



图 7　本文 U-Net 模块详细结构

间的差值之和的相关形式作为损失函数，如 CSPN 采用的损失函数 $L_{org}$，它表示预测值 $\boldsymbol{d}_i$ 和真实值 $\boldsymbol{g}_i$ 之间绝对差值之和的平均值。

$$L_{org}=\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{d}_i-\boldsymbol{g}_i| \tag{3}$$

但这类误差使得场景中远距离目标和近距离目标的单位深度差对损失具有相同的贡献，而随着真实深度 $\boldsymbol{g}_i$ 的增大，绝对估计误差 $|\boldsymbol{d}_i-\boldsymbol{g}_i|$ 也趋于增大，从而使得网络中参数更新受远距离目标的深度误差影响比受近距离目标误差影响大。因此采用此类损失时，网络被训练成更偏向于估计远距离物体的深度[13]。为了解决此问题，采用深度误差的对数 $L_{depth}$ 来作为损失函数：

$$L_{depth}=\frac{1}{n}\sum_{i=1}^{n}\ln(|\boldsymbol{d}_i-\boldsymbol{g}_i|) \tag{4}$$

对于深度图阶梯边缘结构，尽管上述损失函数对深度值方向上的偏移敏感，但对 $x$ 和 $y$ 方向上的偏移以及物体边界的扭曲和模糊相对不敏感。而使用这类损失函数训练神经网络所预测的深度图容易产生边界混合问题的一个主要原因是其对边缘小误差的不敏感。因此，为了使物体边界定位更准清晰，对深度信息求梯度 $L_{grad}$ 来修正边界部分在 $x$ 和 $y$ 方向上的偏移[14]：

$$L_{grad}=\frac{1}{n}\sum_{i-1}^{n}(\ln(|\nabla x(\boldsymbol{d}_i)-\nabla x(\boldsymbol{g}_i)|)+\ln(|\nabla y(\boldsymbol{d}_i)-\nabla y(\boldsymbol{g}_i)|)) \tag{5}$$

其中，$\nabla x(\cdot)$、$\nabla y(\cdot)$ 分别是第 $i$ 个像素在 $x$ 和 $y$ 方向上的导数。但是 $L_{grad}$ 无法对表面高频波动的微小结构误差进行修正，因此采用表面法线损失函数 $L_{normal}$ 来减少物体表面的波动对深度预测的影响[14]：

$$L_{normal}=\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\langle n_i^d,n_i^g\rangle}{\sqrt{\langle n_i^d,n_i^d\rangle}\sqrt{\langle n_i^g,n_i^g\rangle}}\right) \tag{6}$$

其中，$\langle\cdot\rangle$ 表示向量内积：$n_i^d=(-\nabla x(d_i),-\nabla y(d_i),1)^{\mathrm{T}}$，表示预测密集深度图的表面法线，$n_i^g=(-\nabla x(g_i),-\nabla y(g_i),1)^{\mathrm{T}}$ 表示真实深度图的表面法线。最后将 3 种损失函数加权组合在一起形成最终的损失函数 $L$[14]：

$$L=L_{depth}+\lambda L_{grad}+\mu L_{normal} \tag{7}$$

其中，$\lambda$ 和 $\mu$ 为自定义加权系数。

## 3　实验结果及分析

### 3.1　实验数据集与评价指标

本文采用纽约大学公开的室内图像数据集 NYU-Depth-V2[15] 对模型进行训练与测试。NYU-Depth-V2 数据集由微软的 Kinect 深度摄像机拍摄采集，其中包含 464 个场景，原始可用数据约 90 000 张，场景的深度范围为 0～10 m，其中 249 个场景约 50 000 张用于训练，并用官方划分的 654 张图像作为测试。原始成对图像的分辨率为 480×640，考虑到后续 ICSPN 中跳跃连接部分的数据和 2 倍上采样数据尺寸要保持一致，为了保留原始图像更多特征信息，首先下采样至原始分辨率的一半（240×320），然后进行中心裁剪至 224×304 作为网络的输入图像；网络的输入稀疏深度图是通过对数据集中密集深度图进行随机采样获取的点状图，并且稀疏深度图的预处理方式与 RGB 图像的预处理方式相同。

单目深度估计的评价指标主要分为两种[16]，一种为误差指标，本文主要采用均方根误差（root mean squard error，RMSE）和平均相对误差（absolute relative error，REL）；另一种为准确率，本文采用阈值内准确度 $\delta$，其中阈值 $thr$ 越小，说明评价指标越严格。3 个客观指标如式(8)～(10)所示。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{d}_i-\boldsymbol{g}_i)^2} \tag{8}$$

$$REL=\frac{1}{n}\sum_{i=1}^{n}\frac{|\boldsymbol{d}_i-\boldsymbol{g}_i|}{\boldsymbol{g}_i} \tag{9}$$

$$\delta=\max\left(\frac{\boldsymbol{d}_i}{\boldsymbol{g}_i},\frac{\boldsymbol{g}_i}{\boldsymbol{d}_i}\right)<thr,thr=1.25,1.25^2,1.25^3 \tag{10}$$

其中，$\boldsymbol{d}_i$ 表示预测值，$\boldsymbol{g}_i$ 表示真实值，$n$ 表示每张图片中像素的数量。

### 3.2　实验网络参数设置

本文模型的训练与测试是使用深度学习框架 Pytorch 完成的，硬件为单块 NVIDIA RTX 2070S 显卡。使用 SGD 优化器进行优化，设置初始学习率为 0.01，当连续 3 个 epoch 精度没有提升时，学习率将降低到当前学习率的 20%，使用 $10^{-4}$ 的小权重衰减进行正则化。每次训练样本数量 batch_size 设置为 8，总批次 epochs 为 40，稀疏深度图中采样点的数量为 500[11]，占图像像素比例约为 0.7%。

### 3.3　本文损失函数消融实验结果分析

本文在网络结构相同（CSPN 网络结构）的条件下，分

别对损失函数的不同部分进行组合实验,实验结果如表 1 所示,$L_{org}$ 为原始网络 CSPN 采用的损失函数,$L_{depth}$ 作为本文损失函数的主要组成部分,分别与 $L_{grad}$ 和 $L_{normal}$ 进行组合。

**表 1　损失函数消融实验研究**

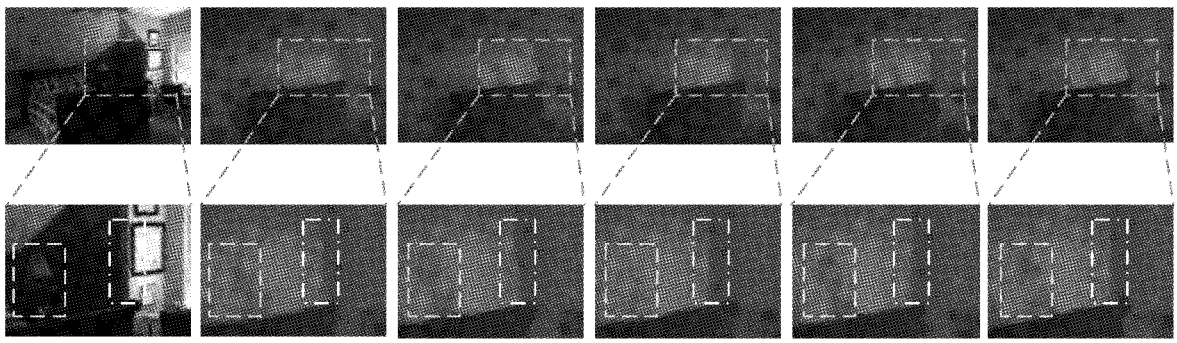
| 方法 | RMSE | REL | $\delta_{1.25}$ | $\delta^2_{1.25}$ | $\delta^3_{1.25}$ |
|---|---|---|---|---|---|
| $L_{org}$ | 0.132 3 | 0.020 3 | 99.02 | 99.80 | 99.96 |
| $L_{depth}$ | 0.142 1 | 0.020 6 | 98.88 | 99.76 | 99.93 |
| $L_{depth}+L_{grad}$ | 0.137 2 | 0.026 6 | 99.09 | 99.82 | 99.96 |
| $L_{depth}+L_{normal}$ | 0.149 5 | 0.029 3 | 98.90 | 99.77 | 99.94 |
| $L$ | **0.128 1** | **0.021 1** | **99.15** | **99.84** | **99.96** |

注:加粗字体为每列最优值。

$L_{depth}$ 与 $L_{depth}+L_{grad}$ 实验结果表明,$L_{grad}$ 的加入降低了实验误差,提高了准确率,这是因为 $L_{grad}$ 能够修正物体边界部分在水平和竖直方向上的偏移。而直接将 $L_{depth}$ 与 $L_{normal}$ 进行组合并不能降低实验误差,相反其组合结果使得误差增大,准确率降低,如 $L_{depth}$ 与 $L_{depth}+L_{normal}$ 实验结果所示,说明在使用 $L_{grad}$ 进行物体边界位置修正前,直接使用 $L_{normal}$ 来修正物体表面高频波动的微小结构误差将对预测结果起到消极作用。最后,将 $L_{depth}$、$L_{grad}$ 和 $L_{normal}$ 三部分进行组合得到 $L$($\lambda$ 和 $\mu$ 取 1),最终预测结果较 $L_{org}$、$L_{depth}$、$L_{depth}+L_{grad}$ 和 $L_{depth}+L_{normal}$ 在均方根误差 RMSE 指标上分别提升了 3.17%、9.85%、6.63% 和 14.31%,验证了本文损失函数各组成部分的必要性和有效性。

### 3.4　整体实验结果对比与分析

图 8 为针对结构变形和边缘混合问题的实验结果放大效果图,其中(a)为 RGB 图像,(b)为原始卷积空间传播网络 CSPN 结果,(c)为本文网络结构改进(记为 ours(net))结果,(d)为本文损失函数改进(记为 ours(loss))结果,(e)为本文网络结构改进与损失函数改进组合形成的 ICSPN 结果,(f)为 RGB 图像对应的真实密集深度图。



(a) RGB图　(b) CSPN　(c) ours(net)　(d) ours(loss)　(e) ICSPN　(f) 真实密集深

图 8　实验效果图结构与边缘对比

从图 8 中虚线框可以看出,CSPN 和 ours(loss)预测的密集深度图中台灯结构变形甚至消失,而 ours(net)和 ICSPN 预测的密集深度图则能够大致显示出台灯结构,说明了网络结构改进中稀疏深度图下采样拼接融合的有效性。从图中点划线框可以看出,CSPN 和 ours(net)预测的密集深度图中物体边缘较为模糊,而 ours(loss)和 ICSPN 中梯度损失的引入能够有效降低物体边缘的位置误差,使之预测的密集深度图中物体边缘部分更为清晰分明。

实验结果的定量比较如表 2 所示,可以看出 CSPN 比 SPN 的均方根误差 RMSE 降低了 18.33%,相对平均误差 REL 降低了 24.81%。而与 CSPN 相比,本文的 ours(net)、ours(loss)和 ICSPN 在均方根误差 RMSE 指标上,分别降低了 10.95%、3.17% 和 17.23%;在平均相对误差 REL 指标上,分别降低了 22.66%、−3.94% 和 28.07%。

**表 2　NYU-V2 数据集实验定量结果对比**

| 方法 | RMSE | REL | $\delta_{1.25}$ | $\delta_{1.25}{}^2$ | $\delta^3_{1.25}$ |
|---|---|---|---|---|---|
| SPN | 0.162 0 | 0.027 0 | 98.50 | 99.70 | 99.90 |
| CSPN | 0.132 3 | 0.020 3 | 99.02 | 99.80 | 99.96 |
| ours(net) | 0.117 8 | 0.015 7 | 99.21 | 99.86 | 99.97 |
| ours(loss) | 0.128 1 | 0.021 1 | 99.15 | 99.84 | 99.96 |
| ICSPN | **0.109 5** | **0.014 6** | **99.33** | **99.89** | **99.98** |

虽然本文最终改进的方法 ICSPN 在准确率指标上相对其他方法提升较小,但预测的密集深度图主观效果明显更接近真实深度图,如图 9 所示。从第 1 行图片的虚线框中桌面与沙发边界部分和第 2 行图片中人体结构部分可以看出,本文方法 ICSPN 较 CSPN 和其他方法更能完整清晰地还原出场景中的结构与边界。

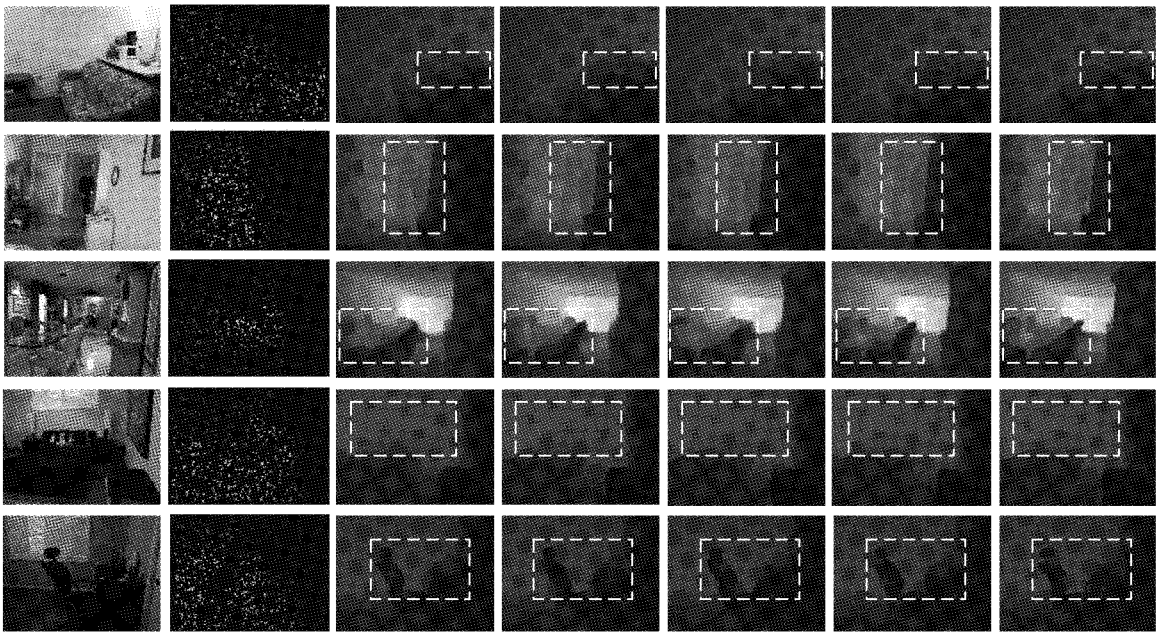



图 9　NYU-V2 数据集实验结果可视化

## 4　结　　论

本文针对 CSPN 网络生成的密集深度图中结构变形及边界混合问题，分别从网络结构与损失函数两个方面进行了改进，并进行了实验验证。

实验结果表明，网络结构改进中，输入稀疏深度图 3 次不同尺寸的下采样能够引导 U-Net 模块中编码器提取不同尺度的特征，同时通过跳跃连接部分传递至解码器的拼接融合结果能够为解码器的上采样部分提供更多信息。U-Net 模块对于稀疏深度图的充分利用减少了密集深度图中物体结构的变形。

损失函数改进中，采用深度误差对数损失函数来替换原始损失函数，由于其对深度值方向上的偏移敏感，对与深度垂直方向上的偏移不敏感，单独使用会使深度图容易产生边界混合问题，因此采用深度信息梯度损失函数修正与深度垂直方向上的偏移。然而前两者对表面高频波动的微小结构误差无法进行修正，而表面法线损失函数则能在一定程度上减少物体表面波动对深度预测的影响，因此最后将 3 种损失函数加权组合形成最终的损失函数，减少了边界混合问题的产生，提高了预测精度。

本文方法的处理速度为 11 fps，即每秒处理的图片数量为 11 张，相对 CSPN 的 12 fps，本文的处理速度并未得到提升。因此，在今后的进一步研究中，将考虑对本文的网络模型进行剪枝，即在尽量不降低模型精度的条件下，将卷积神经网络中冗余的连接或卷积核进行删除，以提高处理速度。

## 参考文献

[1]　徐慧慧. 基于单目图像的深度估计算法研究[D]. 山东：山东大学，2018.

[2]　KUMAR S，KUMAR M. Application of neural network in integration of shape from shading and stereo[J]. Journal of King Saud University-Computer and Information Sciences，2012，24(2)：129-136.

[3]　SKARBEK W. Shape from motion revisited [C]. International Conference on Active Media Technology. Springer，Cham，2014：383-394.

[4]　ZHANG X，LIU Z，JIANG M，et al. Fast and accurate auto-focusing algorithm based on the combination of depth from focus and improved depth from defocus [J]. Optics Express，2014，22(25)：31237-31247.

[5]　张喆韬，万旺根. 基于 LRSDR-Net 的实时单目深度估计[J]. 电子测量技术，2019，42(19)：158-163.

[6]　EIGEN D，PUHRSCH C，FERGUS R. Depth map prediction from a single image using a multi-scale deep network[J]. Neural Information Processing Systems Foundation，2014，3：2366-2374.

[7]　LIU F，SHEN C，LIN G，et al. Learning depth from single monocular images using deep convolutional neural fields [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2015，38(10)：2024-2039.

[8]　FU H，GONG M，WANG C，et al. Deep ordinal

regression network for monocular depth estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2002-2011.

[9] LIU S, DE MELLO S, GU J, et al. Learning affinity via spatial propagation networks[J]. Curran Associates, Inc, 2017: 1519-1529.

[10] CHENG X, WANG P, YANG R. Depth estimation via affinity learned with convolutional spatial propagation network[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 103-119.

[11] MA F, KARAMAN S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image[C]. 2018 IEEE International Conference on Robotics and Automation(ICRA), IEEE, 2018: 4796-4803.

[12] LI A, YUAN Z, LING Y, et al. A multi-scale guided cascade hourglass network for depth completion[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 32-40.

[13] LEE J H, HEO M, KIM K R, et al. Single-image depth estimation based on fourier domain analysis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 330-339.

[14] HU J, OZAY M, ZHANG Y, et al. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries[C]. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019: 1043-1051.

[15] 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.

[16] KHAN F, SALAHUDDIN S, JAVIDNIA H. Deep learning-based monocular depth estimation methods—A state of the art review[J]. Sensors, 2020, 20(8): 2272.

## 作者简介

**刘安旭**,硕士研究生,主要研究方向为图像处理,深度学习。

E-mail: 15950808780m0@sina.cn

**黎向锋**,教授,主要研究方向为计算机智能加工。

E-mail: fxli@nuaa.edu.cn

**刘晋川**,硕士研究生,主要研究方向为图像识别。

E-mail: ljc18761802112@163.com