计算机工程
*Computer Engineering*
ISSN 1000-3428,CN 31-1289/TP

# 《计算机工程》网络首发论文

题目：基于场景模态深度理解网络的单目图像深度理解
作者：陈扬，李大威
DOI：10.19678/j.issn.1000-3428.0059554
网络首发日期：2020-11-23
引用格式：陈扬，李大威．基于场景模态深度理解网络的单目图像深度理解．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0059554

网络首发时间：2020-11-23 17:00:18
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.TP.20201123.1616.001.html



# 基于场景模态深度理解网络的单目图像深度理解

陈扬，李大威*

（东华大学，信息科学与技术学院，上海 201620）

**摘要：**近年来，基于深度卷积神经网络产生的单目深度图像质量已经远高于传统图像处理方法，但目前采用的深度卷积网络存在对无用特征的训练易产生误差积累、基于回归求解连续深度距离预测不理想等问题，导致深度信息不精确、目标边缘模糊、图像细节缺失。本文提出了一种新的深度理解模型，称为场景模态深度理解网络。它基于堆叠沙漏网络反复进行自下而上和自上而下的过程，并在每一层的训练中既使用了离散的深度标签又融合了基本的真实深度图像，同时设计出了可用于进一步改善场景模态预测结果的误差修正子模块和极大似然译码优化子模块用于增强对深度特征的提取效果。实验结果表明本文的方法获得了更加准确的深度信息，在 NYUv2 数据集上的 AbsRel 误差相比 ACAN 方法降低了 0.72%，在 KITTI 数据集上的 SqRel 误差相比 Gan 等人的方法降低了 11.56%；相比 DORN 等方法，本文网络预测出的深度图像中包含了更多的图像细节，并且保持了较好的目标边缘特性。

**关键词：**单目深度估计；场景模态标签；有序回归；误差修正；极大似然译码

开放科学（资源服务）标识码（OSID）：

# Scene Modality Depth Understanding Network for Monocular Depth Estimation

Yang Chen, Dawei Li*

(College of Information Sciences and Technology, Donghua University, Shanghai, 201620 China )

**【Abstract】** In recent years, monocular depth understanding algorithms based on Deep Convolutional Neural Networks (DCNN) have obtained better results than traditional algorithms. However, the problems in currently used DCNN such as error accumulation caused by useless features in the training phase, and undesirable results based on regression methods dealing with the continuous depth estimation task, result in inaccurate depth information, blurred target edges, and missing image details. This paper proposes a new model called Scene Modality Depth Understanding Network for monocular depth estimation. It has a stacked-hourglass framework which can be trained from bottom to top and from top to bottom. Besides, during the training of each layer, both discrete depth labels and continuous depth features of stacked hourglass networks were integrated. Finally, this paper proposes Error Correction Module and Maximum Likelihood Decoding Optimization Module to further improve the results of depth estimation. Experimental results show that the method in this paper obtains more accurate depth information，the AbsRel error on the NYU Depth v2 is reduced by 0.72% compared with ACAN, and the SqRel error on the KITTI data set is reduced by 11.56% compared with Gan *et al.*; compared with methods such as DORN, the depth images estimated in this paper contain more image details and maintain better target edge characteristics.

**【Key words】 monocular depth estimation; scene modality; ordinal regression; error correction; maximum likelihood decoding**

**DOI:** 10.19678/j.issn.1000-3428.0059554

## 1 概述

随着信息技术的发展，人们日益感受到视频场景中深度（距离）信息的重要性，深度图像（Depth Image）是常用于描述场景深度的一种方式。深度图像（也常被称为距离图像）中每个像素值代表着场景中某一点距传感器或扫描仪的距离。目前，深度图像已经被广泛用于无人驾驶[1],[2]、智能机器人[3]、人脸识别[4]等领域。具体而言，在无人驾驶领域，

**基金项目：上海市自然科学基金（No. 20ZR1400800）；国家自然科学基金项目（No. 61603089）**

**作者简介：**陈扬，1994 年生，男，硕士研究生，主要研究方向为计算机视觉、三维重建等；李大威，1986 年生，男，副教授，博士学历，主要研究方向为图像处理、点云分析、机器学习等。**E-mail：daweili@dhu.edu.cn**

车辆的行驶中需要实时获取包含了汽车与周围行人和车辆的距离信息的深度图像。现有的深度图像获取方法如 Kinect、立体匹配、以及激光雷达等，都存在设备昂贵、采集成本高、捕获的深度图像分辨率低以及存在大面积深度缺失等问题，而基于单目彩色图像的深度理解技术以其成本低廉、性能稳定等优势吸引了业界和学界的关注。当前，单目彩色图像的深度理解（估计）技术一般是使用模式识别或机器学习算法从一幅 RGB 图像中估计出其场景中每个像素距传感器的距离。它是计算机视觉领域的一个极具挑战性的任务，主要原因有两个[5]：第一，相机在成像时会不可逆地损失景物三维结构信息，使一张彩色 2D 图像可以与无数个真实场景对应；第二，单幅图像缺乏可被利用来恢复场景深度的有效辅助线索。

早期研究单目彩色图像深度理解的学者大多使用图像中的几何结构特点以及物体与物体之间的相互联系进行计算，例如：从阴影中恢复形状[6]；从对焦[7]或离焦信息[8]获取深度等。此类方法仅适用于有限种类的场景，有些方法还需要额外的辅助信息，严重限制了模型的泛化能力。近些年随着深度卷积网络（Deep Convolutional Neural Network，简称 DCNN）在计算机视觉的多个领域取得突破性进展，一些学者将基于深度学习的方法[9],[10],[11]引入单目彩色图像深度理解中，虽然其产生的深度图像质量远高于传统图像处理方法，但是也有以下三个方面的局限性：1）深度卷积网络从图像中提取大量的特征，然而很多图像特征对深度估计任务没有意义（例如物体颜色、场景的光照或墙壁的纹理和图案等），这些无用特征不仅带来了更大的计算量，还产生了不确定性和学习难度；2）深度学习方法大多将深度理解当作回归问题，这种思路在解决图像分类的时候可能很有效，而深度理解是一个比分类更复杂的连续距离预测问题[12],[13]，用回归求解效果并不理想；3）现存的深度神经元网络伴随着层数的加深易导致错误信息累积[14]，产生低质量的深度估计结果。

受上述三个问题的启发，我们提出了场景模态深度理解网络（Scene Modality Depth Understanding Network，简称 SMDUN）用于解决单目彩色图像深度理解问题。SMDUN 以堆叠沙漏网络为总框架[15]，为避免训练过多的无效特征，同时使用不同分辨率的场景模态离散标签用于指导网络每一层级的特征；为减少由于网络加深而造成的误差积累，我们引入了有序回归码和极大似然译码的思想，进一步优化了对离散标签的学习过程，使 SMDUN 网络总体的多项深度图像估计的定量指标（AbsRel、RMSElog 等）均达到了业界领先水平。我们的贡献总结如下：

1）SMDUN 基于堆叠沙漏网络反复进行自下而上和自上而下的过程，更好地融合了低层次纹理与高级语义特征。在每一层级上我们都用独立的损失函数指导网络丢弃无意义特征。

2）在每一层的训练中既使用了一部分离散的相对深度标签（场景模态），又在堆叠沙漏网络的最后使用连续深度特征，即真实深度标签进行训练，总体上降低了深度理解任务的难度，也因此获取了更好的深度估计效果。

3）用来训练深度神经元网络的离散深度标签通常是 one-hot 型编码的，当估计的深度仅出现小幅波动时，在编码上都会体现出较大的误差。因此，我们引入了有序回归的思想，对用于训练的离散深度标签采用有序回归编码，显著增加了网络的容错能力，提升了训练过程的精确性和稳定性。

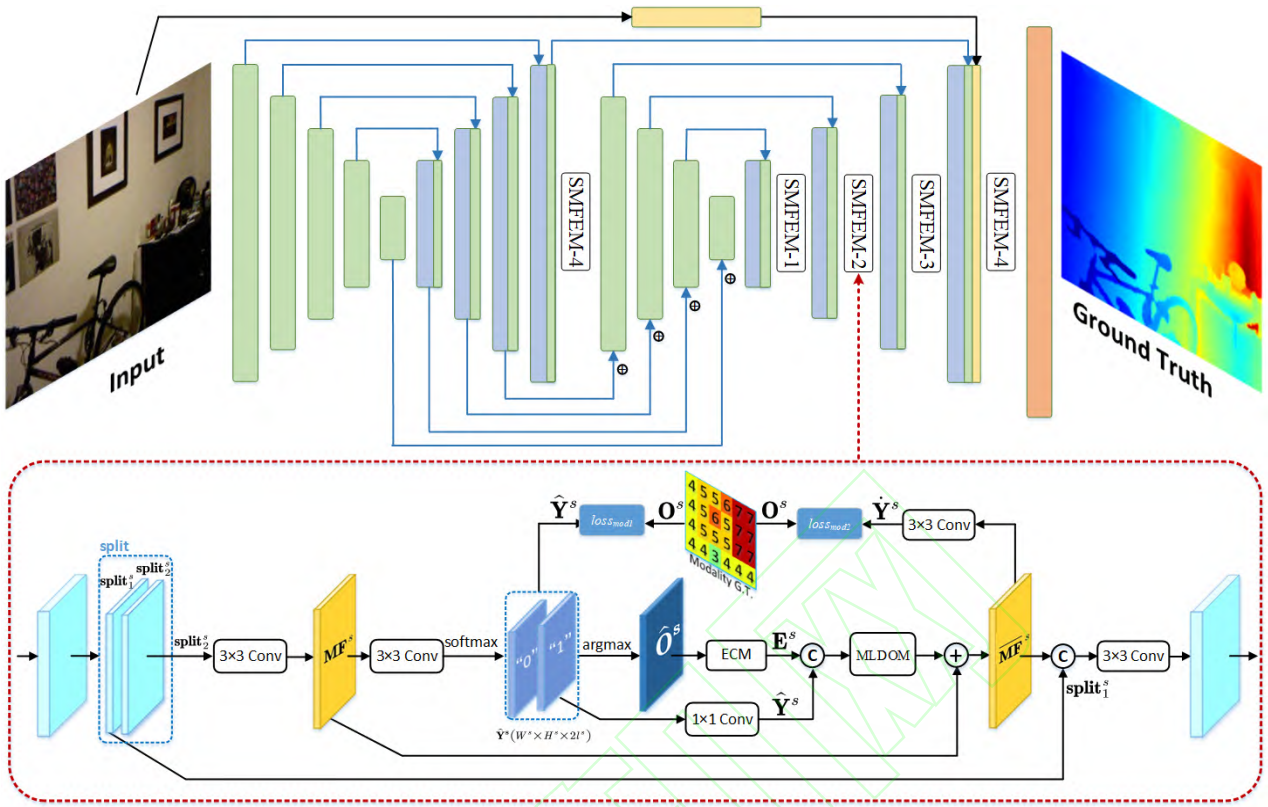


**图 1 场景模态深度理解网络结构**

**Fig. 1 The architecture of Scene Modality Depth Understanding Network**

4）我们引入了极大似然译码的思想[16]，提出了用于进一步改善场景模态预测结果的误差修正模块（Error Correction Module，简称 ECM）和极大似然译码优化模块（Maximum Likelihood Decoding Optimization Module，简称 MLDOM），它们可根据离散的深度结果对网络各层级的特征进行优化。我们认为接收码是网络预测出的有序回归码，而发送码是正确的有序回归标签，所设计的极大似然译码优化模块不仅能够进一步减少有序回归的错误，同时还以贝叶斯后验概率的方式，使接受码尽量逼近发送码，产生更强的深度理解能力。

5）我们把 SMDUN 在 NYUv2 和 KITTI 数据集上分别进行了定性与定量实验，其结果优于当前的主流 DCNN 算法。

## 2 相关工作

从单幅彩色图像中理解深度是一个极具挑战性的任务。早期研究工作大多使用图像中的几何结构特点，因此仅适用于有限种类的场景，模型的泛化能力比较有限。现如今，深度卷积网络在计算机视觉的多个领域取得突破性进展，基于深度学习的方法已经成为研究单目彩色图像深度理解领域的主流。相比早期方法使用人工定义的特征进行深度理解的研究，基于深度卷积网络的方法能从彩色图像中提取更多有利于深度理解的线索，因此得到的预测深度图像质量更佳。Eigen 等人[9]的工作是使用深度学习进行图像深度理解的开创性成果，他们提出了一个双栈卷积神经网络，网络首先得到一个粗糙的全局预测结果，再使用局部特征对其进行优化。文献[17]提出了一个双流 CNN 从单幅图像中恢复深度，双流网络中的一条流产生深度特征而另外一流产生深度梯度特征，最后融合二者信息产生更精细的深度图。文献[18]利用深度学习网络中间层的输出来提供互补信息，通过连续 CRF 模型对模型中间层的输出信息进行整合，实现了对单幅图像的有效深度估计。文献[19]提出了一种无监督的单目彩色图像深度理解框架，他们使用立体图对基于光度重建损失函数以进行视差估计，从而得到深度图像。在此基础上，文献[20]进一步提出了左右一致性检验，

并结合 L1 损失和结构相似性(structural similarity index，简称 SSIM) 作为损失，得到更平滑的深度图预测结果。文献[21]基于特定的几何结构，提出了一种基于几何感知的对称域自适应框架，通过训练两个图像样式转换器和深度估计器，实现彩色图像与深度图像的样式转换。上述提及的深度学习方法大多将深度理解当作回归问题来解决，这类方法在解决图像分类和语义分割问题的时候可能很有效，然而深度理解任务中每一像素的深度都是一个连续的值，对其的预测远比离散的分类问题复杂。基于此，文献[13]基于有序回归的思想，将连续的深度估计任务转换为带有前后关联性的离散深度标签分类问题，降低了深度理解的难度。但此类方法把深度图像首先离散化作为训练标签已丢失了大量的深度信息，因此得到的预测图像也存在一定程度的特征丢失。文献[22]基于图像级全局特征和像素级局部信息的特征，通过有序回归的概率信息将离散的有序回归结果转换为连续值进行处理，但此类方法仅仅通过分类概率推测出一个连续的深度值，标签在离散化阶段丢失的信息并没有从本质上得到弥补。

此外，上述提及的方法都是基于图像的纹理信息进行深度理解，易使网络学习到大量不相关的特征（例如墙壁的纹理特征），这些特征对从图像中提取深度信息没有任何作用，反而带来了更大的计算量，产生了不确定性，甚至带来了学习难度。因此，文献[15]将不同层级的特征反复处理和融合以提取有效特征。文献[23]通过在每一层级单独计算损失函数而丢弃无用特征。但上述两种机制随着层数的增加易产生误差积累，在预测深度图中产生不合理的几何分布。文献[14]认为网络需要特征优化机制，并提出了基于网络先前层级的特征对当前层级特征进行补充和修正的策略，然而在这种特征优化机制下，先前低层级包含的特征远没有当前层级的特征丰富，因此特征的优化能力有限。

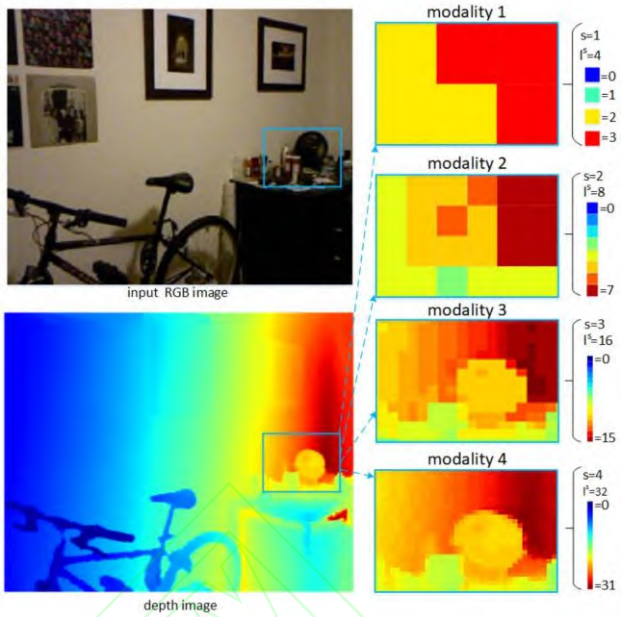


**图 2 深度图与场景模态标签**

**Fig.2 Depth image and scene modality labels**

## 3 场景模态深度理解网络

这一节我们将详细介绍场景模态深度理解网络（Scene Modality Depth Understanding Network，SMDUN），网络结构如图 1 所示。该网络具有如下特点：首先，SMDUN 总体基于堆叠沙漏结构，同时使用连续与离散两种标签进行训练；其次，我们提出场景模态特征提取模块 SMFEM (Scene Modality Feature Extraction Module)，在 SMDUN 的多层次堆叠沙漏结构中逐次使用场景模态特征提取，基于一个综合损失函数指导网络从低层级到高层级理解深度信息；第三，我们提出误差纠正模块（Error Correction Module，简称 ECM）和极大似然译码优化模块（Maximum Likelihood Decoding Optimization Module，MLDOM）用于修正中间层级的错误特征，减少误差累计。

### 3.1 SMDUN 的网络框架

文献[15]使用了堆叠沙漏结构的网络解决了图像中人体关节点检测的问题，文献[24]证明了堆叠沙漏网络能够对双目立体视觉系统进行深度估计。因此，我们的 SMDUN 以堆叠沙漏网络为基础，致力于图像深度特征的提取和理解。网络通过中间指导和反复自下而上与自上而下的过程，有效地融合了低层次纹理与高级语义特征。SMDUN 中的双沙漏结构如图 1 所示，RGB 图像经第一个沙漏网络的编

码器提取图像底层特征，特征图的分辨率从$W \times H$逐渐降低至$W/32 \times H/32$；在解码器中通过跳链补充图像底层特征，并逐层级提高特征的分辨率至$W/2 \times H/2$；第二个编码器降低特征的分辨率至$W/32 \times H/32$并将每一层级输出特征与先前解码器相应层级的特征相加；最后特征在第二个解码中通过跳链从前一个编码器和彩色图像中补充图像底层特征，并输出分辨率为$W \times H$的深度估计结果。

SMDUN 采用的是逐层级优化的思想（stage-wise refinement）以降低网络的不确定性和无效特征的影响，进一步提高网络的收敛能力和预测精度。大多数深度理解网络中（例如文献[9],[14]）使用连续的深度值标签指导网络的中间层级特征，易带来不确定性，导致网络难以学到有效特征。文献[13]将有序回归的思想引入深度理解与估计任务中，将连续的深度估计任务转换为带有前后关联性的离散深度标签分类问题，降低了深度理解的难度。同时，有序回归也成为了架设在深度特征理解与编码译码方法之间的桥梁，对深度估计问题引入了坚实的编译码理论的支撑。

虽然离散图像标签具有降低计算量的作用，但与连续的深度真实图相比，存在一定程度的信息丢失。此外，对有序回归码的特征提取与计算中还容易产生两类错误。因此，我们提出场景模态特征提取模块 SMFEM 模块以解决上述两个问题。如图 1 所示，我们添加 SMFEM 模块于沙漏网络解码器的各阶段输出后，达到逐层级优化的效果。在 SMFEM 中，我们将输入特征图分为两个部分$\mathbf{split}_1^s$和$\mathbf{split}_2^s$，我们保留$\mathbf{split}_1^s$作为前馈残差的低层次特征，对$\mathbf{split}_2^s$使用场景模态离散标签训练并经过 MLDOM 模块优化得到特征$\overline{\mathbf{MF}}^s$。随后，我们将$\mathbf{split}_1^s$和$\overline{\mathbf{MF}}^s$两部分特征进行拼接，确保所获得特征的完整性。最后，通过 3×3 卷积得到 SMFEM 的输出特征。

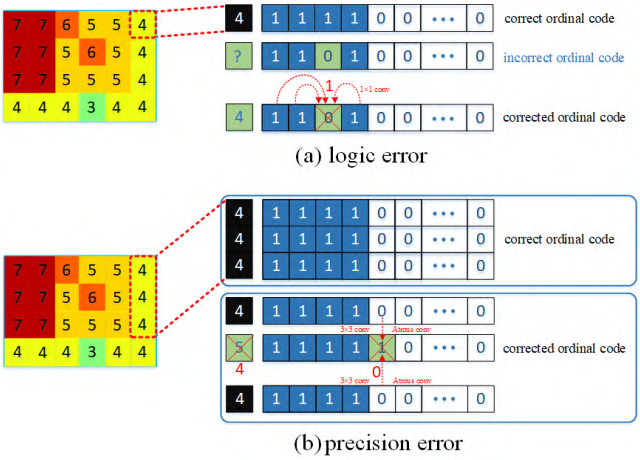


**图 3 有序回归码的两种错误例子与误差纠正模块中对应的卷积修正方式展示。(a)展示了有序回归码的逻辑错误和使用 1X1 卷积修正的方式; (b)展示了有序回归码的精度错误和使用 3X3 卷积与空洞卷积对其修正的方式。**

**Fig.3 Two error examples of ordinal regression codes and the corresponding convolution correction methods in the error correction module are displayed. (a) shows the logic error of the ordinal regression code and the correction of using 1X1 convolution; (b) shows the precision error of the ordinal regression code and the correction of using 3X3 convolution and dilated convolution.**

### 3.2 场景模态特征的提取

逐层级的特征优化被广泛用于结构化的训练任务中，例如图像语义分割、深度估计和边缘检测[25-27]。文献[23]通过多分辨率的训练标签指导特征，并在每一层级之后计算独立的损失函数，这样做能有效减少无用特征的学习。因此，业界很自然地想到多次使用不同采样率进行降采样后的真实深度图像作为多分辨率的训练标签进行训练[9,14]。然而，真实深度图像中每个位置的值是连续的浮点值，增加了训练的难度和不确定性。实际上，深度图像中最重要的信息是远与近的相对概念，我们完全可以使用离散的数字类别标签对相对距离进行编码，然后使用离散编码后的深度图进行训练。此时离散的标签类似于语义分割中物体类的概念，可以参照成熟的语义分割网络的标签训练方式进行训练。我们把基于相对距离关系进行离散化后的真实深度图像称为“场景模态”。为了能够针对场景模态标签进行训练，我们设计了场景模态特征提取模块 SMFEM，其具体结构可见图 1 的下半部分。

针对上述问题，我们首先提出了多分辨率的场景模态标签的构建方式。我们从深度图像中提取 $M$ 种场景模态标签，如图 2 所示，多分辨率的场景模

态标签表示为 $\mathbf{Modality}=\{\mathbf{M}^s, s=1,2,\ldots,M\}$，其中 $\mathbf{M}^s$ 为 SMDUN 中第 $s$ 种标签，$\mathbf{M}^s(x,y)$ 表示标签 $\mathbf{M}^s$ 中位置 $(x,y)$ 的值，$W^s$ 和 $H^s$ 为标签宽度和高度，$\mathbf{M}^s$ $x,y$ 标签的取值区间为 $0,1,\cdots,l^s-1$，$l^s$ 为本级场景模态可表示的相对距离级数，该值我们在之后所有的计算中取 2 的幂次。

场景模态标签由相对距离计算生成，相当于通过远、进、较远与较近等模糊概念描述图像的空间分布。计算相对距离场景模态标签步骤如下：

首先通过公式（1）的线性归一化算法计算得到每一个位置的相对距离深度标签 $\mathbf{D}_r$：

$$\mathbf{D}_r(x,y)=\frac{\mathbf{Depth}(x,y)-D_{\min}}{D_{\max}-D_{\min}} \tag{1}$$

其中，$\mathbf{Depth}$ 表示当前深度图，$D_{\min}$ 表示当前深度图中最小的深度值，$D_{\max}$ 表示当前深度图的最大深度值。

随后对相对距离标签进行非均匀离散化得到离散的相对距离标签为 $\mathbf{D}_d$，离散化阈值为 $t_i^s\in\{t_0^s,t_1^s,\cdots,t_{l^s}^s\}$，离散化阈值由公式（2）计算得到，

$$t_i^s=\exp\left(\log(\alpha)+i\cdot\log(\beta/\alpha)/l^s\right) \tag{2}$$

其中，$\alpha$ 表示本层场景模态中全部 $\mathbf{D}_r$ 标签的最小值，$\beta$ 为 $\mathbf{D}_r$ 标签最大值，$l^s$ 表示离散化区间数。为了避免实际距离为 0 的情况，我们对 $\alpha$ 和 $\beta$ 添加偏移量 1 成为 $\alpha^*$ 和 $\beta^*$，以避免无法计算 log 的问题。因此实际使用的非均匀离散化取值区间为 $[\alpha^*,\beta^*]$。

接下来在 $\mathbf{D}_d$ 中均匀划出 $W^s\times H^s$ 个区域（$W^s$ 与 $H^s$ 的取值与这一级的场景模态标签的长宽相关），并计算每一区域的平均值，得到粗糙的场景模态标签。

最后，对粗糙的场景模态标签重复公式（1）计算相对距离的过程和公式（2）离散化的过程，得到最终的场景模态标签 $\mathbf{M}^s$。

标签 $\mathbf{M}^s$ 由数字 0 到 $l^s-1$ 构成，$\mathbf{M}^s$ 与阈值 $t_i^s$ 的关系如公式（3）所示：

$$\mathbf{M}^s(x,y)=i,\quad s.t.\quad t_i^s<\mathbf{Depth}(x,y)\le t_{i+1}^s \tag{3}$$

为了增加网络的容错能力，提升训练过程的稳定性，我们对场景模态离散标签并没有使用常见的 one-hot 型编码。比如对某个位置的真实相对深度为 4，网络预测为 5，对于 one-hot 型编码而言，其错误产生的损失与预测为 8 所产生的损失是差不多的。然而实际上相对深度具有一定的关联性，5 与 4 相比于 8 与 4 之间更加接近，给予一个更小的损失比较合理。所以我们设计了有序回归码的网络。

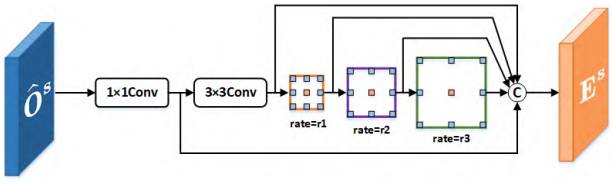


**图 4 误差纠正模块结构图**

**Fig.4 The Error Correction Module**

有序回归方法实际上是将一个复杂的多分类任务用 $l^s$ 个简单的二分类任务代替，在网络的训练和推理过程中，我们将标签 $\mathbf{M}^s$ 转换为有序回归码 $\mathbf{O}^s$，$\mathbf{O}^s$ 的分辨率为 $W^s\times H^s\times L^s$，其中 $L^s=l^s-1$。$\mathbf{M}^s$ 与 $\mathbf{O}^s$ 在 $(x,y)$ 位置满足(4)中的关系：

$$\begin{cases}\mathbf{O}^s(x,y,0:\mathbf{M}^s(x,y)-1)=1\\ \mathbf{O}^s(x,y,\mathbf{M}^s(x,y):L^s-1)=0\\ \mathbf{M}^s(x,y)=\sum_{i=0}^{L^s-1}\mathbf{O}^s(x,y,i)\end{cases} \tag{4}$$

$\mathbf{O}^s$ 码在 $x,y$ 的每一维度值实际上是一个二分类任务，其值是 0 与 1 的训练过程，在训练中我们希望得到一个二分类标签的概率张量 $\mathbf{Y}^s$，其分辨率为 $W^s\times H^s\times 2\times L^s$。$\mathbf{Y}^s$ 由两层 $W^s\times H^s\times L^s$ 大小的特征层构成（在图 1 中以“0”与“1”标示），其中“0”特征层表示经过网络得到的有序回归码中二分类结果为标签 0 的概率，“1”是把有序回归码每位为 1 的概率按顺序排序后得到的特征层。

在网络的推理阶段，从预测的有序回归结果 $\mathbf{Y}^s$ 可以通过 argmax 函数得到 $\hat{\mathbf{O}}^s$，从而估计出 $\hat{\mathbf{M}}^s$，这一过程如公式(5)、(6)所示，

$$\hat{\mathbf{O}}^s(x,y,i)=\eta(\mathbf{P}^s(\text{"1"})>\mathbf{P}^s(\text{"0"})) \tag{5}$$

其中，$\eta(\cdot)$ 为指示函数，满足 $\eta(true)=1$ 而 $\eta(false)=1$。$\mathbf{P}^s(\text{"0"})$ 表示位置为 $(x,y)$ 处的有序回归码的第 i 位为 0 的概率，而 $\mathbf{P}^s(\text{"1"})$ 表示码的第 i 位为 1 的概率，$\mathbf{P}^s(\text{"0"})$ 和 $\mathbf{P}^s(\text{"1"})$ 中的同一位置的值加起来和为 1。满足如下公式：

$$\mathbf{P}^s\ "\mathbf{0}" = \frac{\exp\ \hat{\mathbf{Y}}^s\ "\mathbf{0}"}{\exp\ \hat{\mathbf{Y}}^s\ "\mathbf{0}"\ + \exp\ \hat{\mathbf{Y}}^s\ "1"} \tag{6}$$

$$\mathbf{P}^s\ "1" = 1 - \mathbf{P}^s\ "\mathbf{0}" \tag{7}$$

而$\mathbf{M}^s\ x,y$ 与$\mathbf{O}^s\ x,y,i$ 满足(8)。

$$\mathbf{M}^s\ x,y = \sum_{i=0}^{t^s-1} \mathbf{O}^s\ x,y,i \tag{8}$$

由场景模态再得到相对深度值如公式(9)所示:

$$\mathbf{D}_r^s\ x,y = \frac{t^s_{\mathbf{M}^s\ x,y} + t^s_{\mathbf{M}^s\ x,y+1}}{2} - 1 \tag{9}$$

深度卷积网络存在欠拟合和过拟合现象，多层级的深度卷积网络在训练和推理阶段将当前层级的结果直接送入下一层模块，也将当前层的误差和噪声传递到后续网络中，导致误差积累并最终反映在预测深度图中。因此，我们需要对网络中的一些错误进行及时校正。在估计出的场景模态有序回归码$\hat{\mathbf{O}}^s$中，不可避免地包含两种类型的错误。第一种是有序回归码的内在逻辑错误，第二种是有序回归的二分类精度错误。两种错误的范例如图 3 所示。

图 3(a)展示了估计出的某一层场景模态一个位置上出现了逻辑错误，具体表现为在值为 4 的场景模态上，其本应是“1,1,1,1”的有序回归码在第 2 位（从第 0 位开始）发生了错误，变为了“1,1,0,1”。从逻辑上而言这是错误的，因为我们定义的有序回归码不能出现 0,1 交替的情况。然而在网络训练的过程中，难以避免这类错误，而且我们无法在训练中直接对有序回归码的具体值进行赋值操作，将错误的 0 替换为 1，只能通过以卷积和反向传播的形式进行纠错。图 3(b)展示了场景模态某一位置上出现了精度错误，具体表现为在值为 4 的场景模态上，其本应是“1,1,1,1”的有序回归码变为了“1,1,1,1,1”，导致该位置场景模态实际上变为了 5，将影响后续深度估计的精确性（因为场景模态反映相对距离）。为了解决这两种训练中常见的有序回归码错误。我们设计了一个包含几种基本卷积的误差纠正模块（Error Correction Module，简称 ECM）。

ECM 模块的内部结构如图 4 所示。$\hat{\mathbf{O}}^s$中出现的两种有序回归错误，本质上都是在一系列二分类任务上产生了错误分类。针对第一种内在逻辑错误，我们通过 1×1 卷积学习有序回归码的规则。如图 3(a)所示，经过 1x1 卷积后，同一串码前后的正确码字经过卷积能对逻辑错误位产生影响，从一定程度上消除错误；针对第二种精度错误，只凭借当前场景模态位置的信息不足以修正，我们通过 3x3 卷积和多层空洞卷积，以类似于[28]的多层空洞池化模块的卷积连接方式（如图 3(b)所示），充分提取场景模态中相邻位置的特征以克服当前的分类精度错误。在 ECM 模块的最后，将每一阶段产生的多尺度特征进行拼接，得到修正后的场景模态特征$\mathbf{E}^s$。

### 3.3 极大似然译码优化模块

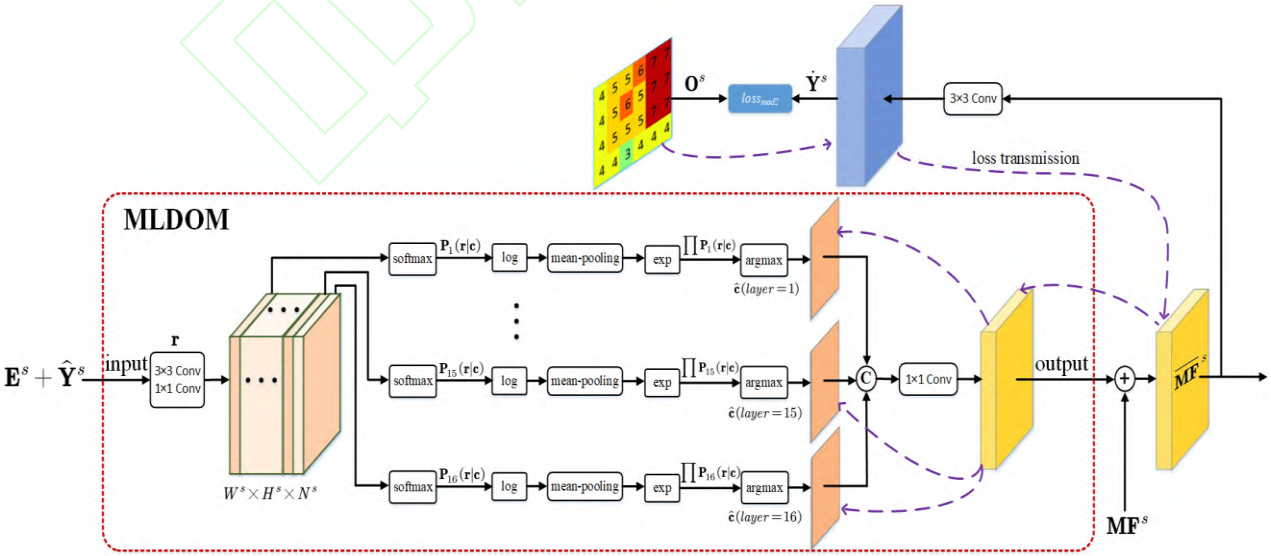


**图 5 极大似然译码优化模块**

**Fig.5 The Maximum Likelihood Decoding Optimization Module**

这一节，我们设计了一个极大似然译码优化模块（Maximum Likelihood Decoding Optimization Module，简称 MLDOM）实现极大似然译码的思想。该模块把预测的有序回归码 $\hat{\mathbf{O}}^s$ 当作是包含错误和噪声的接收码，把场景模态真实值的有序回归码 $\mathbf{O}^s$ 当作发送码，使接收码能最大限度地逼近发送码。该模块 MLDOM 从该层级的预测结果 $\hat{\mathbf{O}}^s$ 中得到场景模态的优化特征，并与 $\mathbf{MF}^s$ 相加得到优化后的场景模态特征 $\overline{\mathbf{MF}}^s$，最后将 $\overline{\mathbf{MF}}^s$ 与 $\mathbf{split}_1^s$ 拼接得到整个 SMFEM 模块的输出特征。

在信息论的译码任务中，发送码可表示为 $\mathbf{c}=[c_1,c_2,\cdots,c_N]\in\mathbf{X}^N$，接收码为 $\mathbf{r}=[r_1,r_2,\cdots,r_N]$。其中 $\mathbf{c}$ 表示 $N$ 个 $q$ 进制码元组成的码字，$\mathbf{X}=[x_1,x_2,\cdots,x_q]$ 为字符集。极大似然译码从可测的发送码 $\mathbf{c}$ 的 $q^N$ 种可能性中找到后验概率最大的发送码 $\hat{\mathbf{c}}$，标准最大后验概率译码的过程如公式(10)所示：

$$\hat{\mathbf{c}}=\arg\max_{\mathbf{c}} p(\mathbf{c}|\mathbf{r}) \tag{10}$$

然而，在实际的物理系统中，只存在信息从发到收的因果前向转移概率（即先验概率）$p(\mathbf{r}|\mathbf{c})$，信道中并不存在后验概率 $p(\mathbf{c}|\mathbf{r})$。因此只能尝试通过先验概率近似计算后验概率。通过贝叶斯公式表

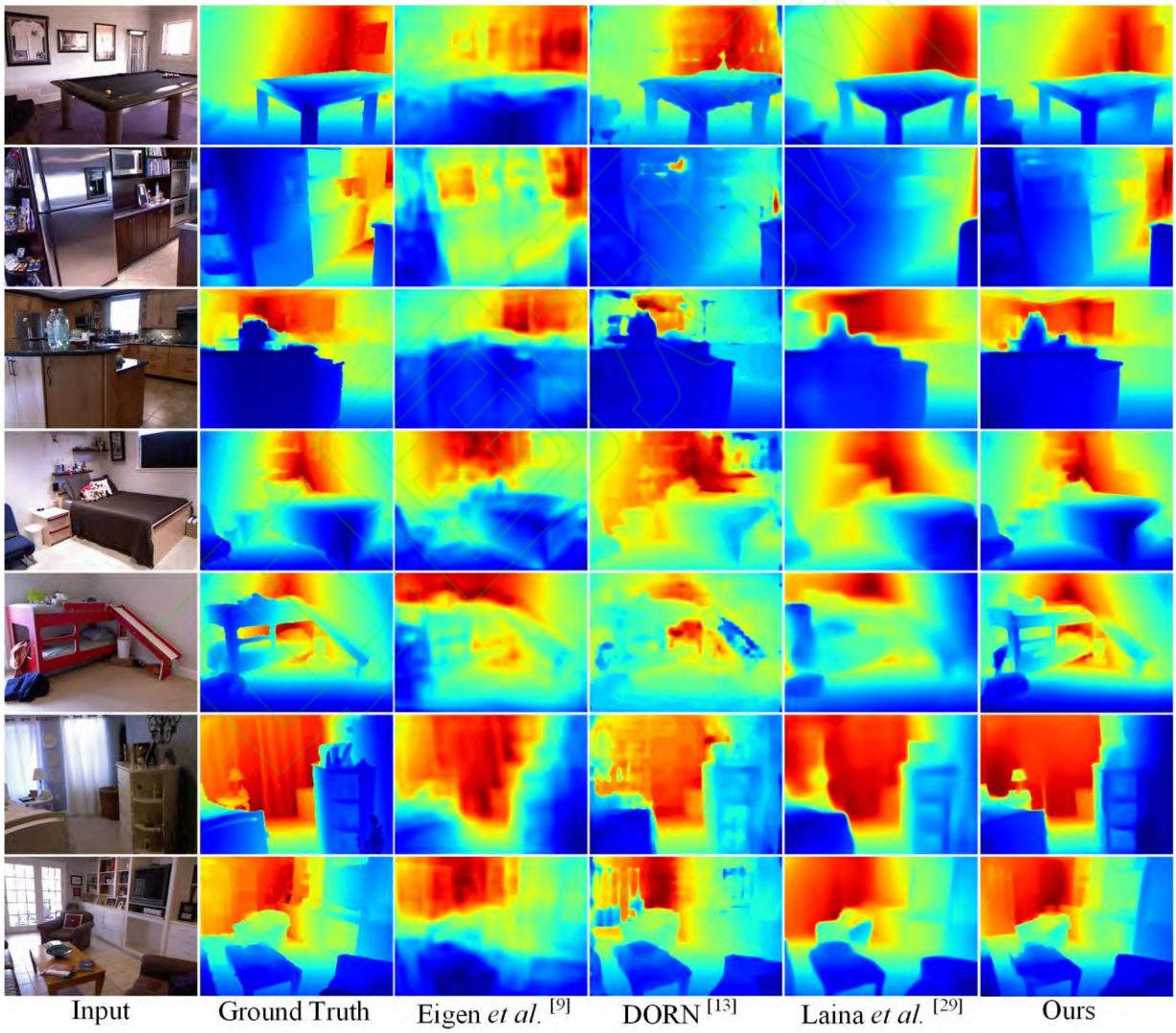


**图 6. SMDUN 网络与最新算法在 NYUv2 数据集的定性结果对比。这里以行的方式列举了 7 个场景的定性对比结果。相比于其他的几种网络，我们的网络最接近于第二列展示的深度真实值。**

**Fig.6 Visual comparisons on NYU depth V2 between some baselines and our proposed method. Here are the qualitative comparison results of 7 scenes in a row. Compared to several other networks, our results are closest to the ground truth shown in the second column visually.**

示的两个概率的关系如下：

$$p(\mathbf{c}|\mathbf{r}) = \frac{p(\mathbf{c})p(\mathbf{r}|\mathbf{c})}{p(\mathbf{r})} \tag{11}$$

由于发送码和接收码都可以是$q^N$种可能性中的一种，假设每种码的概率相同，则有$p(\mathbf{c}) = p(\mathbf{r}) = 1/q^N$，此时后验概率与先验概率同时达到最大，公式(10)可以转化为下式：

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} p(\mathbf{r}|\mathbf{c}) = \arg\max_{\mathbf{c}} \prod_{j=1}^{N} p(\mathbf{r}_j|\mathbf{c}_j) \tag{12}$$

若要在大小为$W^s \times H^s$的场景模态层上计算极大似然译码，则需要计算如下公式：

$$\mathbf{C} = \arg\max_{\mathbf{c}} \prod_{\substack{x=1,\\ y=1}}^{W^s,H^s} \prod_{j=1}^{N} p(\mathbf{r}_j(x,y)|\mathbf{c}_j(x,y)) \tag{13}$$

此时不仅计算量过大，而且似然函数也难以确定。因此我们以卷积的方式实现局部的极大似然译码，寻求以较低的计算量得到一个次优解。用局部的计算来近似极大似然译码是具有充分理由的，首先：场景模态为类似于深度图像的相对距离，其中每个坐标的相对深度其实与邻域存在比较紧密的联系，因为目标的表面往往是深度连续的[24]。其次：对于图像而言，目标级信息描述了图像场景的整体结构和具体物体的粗略位置关系，即全局特征；而像素级信息精确了物体表面在场景中的深度值，主要体现为局部的特征，完全可以通过网络在训练阶段以卷积和池化学习到。在不同图像中场景会发生改变，但场景中同类的物体特征却不会变。例如，在客厅学习到的桌子的特征同样适用于在厨房中的桌子。因此我们认为局部特征不会随着场景的改变而失效，具有较高的鲁棒性。

基于上述分析，首先将输入 MLDOM 模块的特征转化为$W^s \times H^s \times N^s$的特征层，其中$N^s$为当前极大似然译码相关的码长，然后我们将极大似然译码转换为局部最优的译码过程。这个过程主要包含两步：第一，我们把特征层平均划分为 16 层的子特征层，其中每层通道数为$N^s/16$，然后分别进行极大似然译码计算；第二，在每个子特征层中，通过 log 似然的方式将概率连乘变为连加，然后通过 5×5 的平均池化将连加限制在局部范围内进行，最后通过 argmax 获取局部最优的特征编码，近似完成公式(13)的计算。图 5 是极大似然译码优化模块的内部结构，其中对译码过程的先验概率逼近也进行了展示，由场景模态标签带来的损失通过紫色虚线传导至了每一个单独的译码过程，保证了最优译码方向的正确

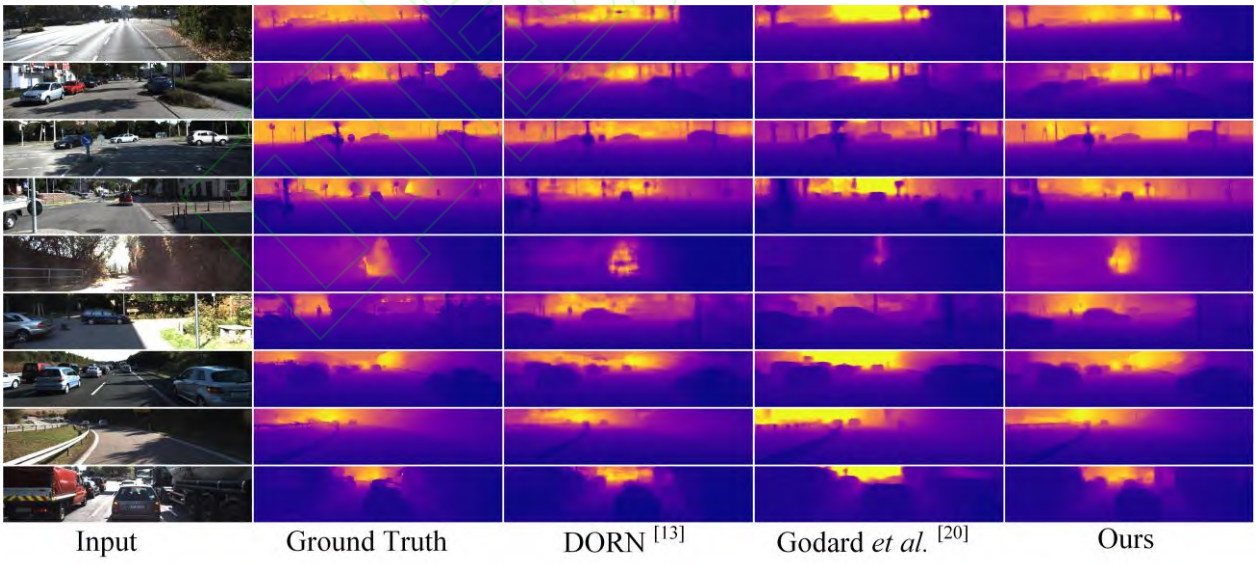


**图 7. SMDUN 网络与最新算法在 KITTI 数据集的定性结果对比。这里以行的方式列举了 9 个街景的定性对比结果。相比于其他的几种网络，我们的网络最接近于第二列展示的深度真实值。**

**Fig.7 Some examples of depth prediction result on KITTI between some baselines and our proposed method. Here are the qualitative comparison results of 9 scenes in a row. Compared to several other networks, our results are closest to the ground truth shown in the second column visually.**

性。

### 3.4 损失函数

这一小节我们定义了所提出的 SMDUN 的损失函数 $loss_{total}$，如公式(14)所示：

$$loss_{total} = \alpha_{im} \cdot loss_{img} + \alpha_{mod} \cdot loss_{mod} \tag{14}$$

总损失函数主要由两部分构成，第一部分为预测得到的深度图像和真实深度图像标签之间的损失 $loss_{img}$，第二部分为场景模态标签的损失 $loss_{mod}$。第一部分损失在整个堆叠沙漏网络的最后进行计算，而第二部分的损失在每一层 SMFEM 内进行计算。

#### 3.4.1 预测深度图和标签深度图之间的损失

深度图像预测误差 $loss_{img}$ 主要由 Inverse-Huber 损失[29]和 SSIM 指标[10]两部分组成，如公式(15)所示：

$$loss_{img} = \alpha \cdot \mathrm{B}\left(\mathbf{D} - \hat{\mathbf{D}}\right) + \left(1 - \alpha\right) \cdot \frac{1 - \mathrm{SSIM}\left(\mathbf{D}, \hat{\mathbf{D}}\right)}{2} \tag{15}$$

其中，$\hat{\mathbf{D}}$ 为网络的预测结果，$\mathbf{D}$ 为深度图像标签，$\mathrm{B}(\cdot)$ 为 Inverse-Huber 损失，如公式(16)所示，SSIM 函数用以计算两幅图像间的相似度，$c$ 为阈值。

$$\mathrm{B}(x) = \begin{cases} |x|, & |x| \le c \\ \dfrac{x^2 + c^2}{2 \cdot c}, & |x| > c \end{cases} \tag{16}$$

#### 3.4.2 场景模态损失

我们定义场景模态标签损失 $loss_{mod}$ 为全部 SMFEM 模块的有序回归损失之和，如公式(17)所示：

$$loss_{mod} = \sum_{s=1}^{M} loss_{mod}^{s} \tag{17}$$

其中，M 为 SMFEM 模块的个数，$loss_{mod}^{s}$ 为第 s 个的 SMFEM 的损失，该损失由特征提取和特征优化两部分构成，如公式(18)所示：

$$loss_{mod}^{s} = \beta_s \cdot loss_{mod1}^{s}\left(\hat{\mathrm{Y}}^{s}, \mathrm{O}^{s}\right) + \gamma_s \cdot loss_{mod2}^{s}\left(\dot{\mathrm{Y}}^{s}, \mathrm{O}^{s}\right) \tag{18}$$

其中，$\mathbf{O}^{s}$ 为第 s 层的场景模态的有序回归码标签，$\hat{\mathrm{Y}}^{s}$ 为场景模态的预测结果，$\dot{\mathrm{Y}}^{s}$ 为优化和修正后的场景模态预测结果，$\beta_s$ 为约束第一项的权重，$\gamma_s$ 为约束优化场景模态输出的权重。函数 $loss_{mod1}$ 与 $loss_{mod2}$ 计算方式相同，$loss_{mod1}$ 的定义见公式(19)：

$$loss_{mod1}^{s}\left(\hat{\mathrm{Y}}^{s}, \mathrm{O}^{s}\right) = -\frac{1}{W^{s} \times H^{s}}\left(\sum_{k=0}^{\mathbf{M}^{s}_{x,y}-1} \log \mathbf{P}^{s}(\text{"1"}) + \sum_{k=\mathbf{M}^{s}_{x,y}}^{x^{s}-1} \log \mathbf{P}^{s}(\text{"0"})\right) \tag{19}$$

## 4 实验结果与分析

在这一节我们通过实验验证 SMDUN 的深度理解有效性。通过设计不同的剥离实验分析网络各个部分的有效性，并将我们的网络与当前流行的其他先进方法进行了定性和定量比较。

**表 1 剥离实验的定量对比。其中第一行指标边的向上箭头表示该指标越高越好，向下箭头表示该指标越低越好。表中最佳值以加粗表示。**

**Table1 Ablation studies. The upward arrow on the indicator side of the first row indicates that the higher the indicator is, the better, and the downward arrow indicates that the lower the indicator is, the better. The best value in the table is expressed in bold.**

| Type | Mlog10E↓ | AbsRel↓ | RMSE↓ | δ1↑ | δ2↑ | δ3↑ |
|---|---|---|---|---|---|---|
| 1 | 0.062 | 0.149 | 0.630 | 0.801 | 0.954 | 0.987 |
| 2 | 0.060 | 0.143 | 0.607 | 0.815 | 0.957 | 0.988 |
| 3 | 0.061 | 0.144 | 0.612 | 0.803 | 0.952 | 0.986 |
| SMDUN | **0.058** | **0.137** | **0.585** | **0.820** | **0.960** | **0.989** |

**表 2 多种方法在 NYUv2 数据集上的定量对比结果。表中最佳值以加粗表示，第二好的值以下划线突出。**

**Table 2 Quantitative comparison results of multiple methods on NYU Depth v2. The best value in the table is expressed in bold, and the second best value is highlighted underlined.**

| Method | Mlog10E↓ | AbsRel↓ | RMSE↓ | δ1↑ | δ2 | δ3↑ |
|---|---|---|---|---|---|---|
| Eigen *et al.* [9] | - | 0.214 | 0.877 | 0.611 | 0.880 | 0.971 |
| DCNF [11] | - | 0.230 | 0.824 | 0.614 | 0.883 | 0.971 |
| Joint HCRF [33] | - | 0.220 | 0.745 | 0.605 | 0.890 | 0.970 |
| Li *et al.* [17] | 0.063 | 0.143 | 0.635 | 0.788 | 0.958 | <u>0.991</u> |
| Cao *et al.* [12] | <u>0.060</u> | 0.141 | 0.540 | 0.819 | **0.965** | **0.992** |
| Li *et al.* [34] | **0.058** | 0.139 | <u>0.505</u> | 0.820 | 0.960 | 0.989 |
| PHN [14] | - | 0.144 | **0.501** | **0.835** | <u>0.962</u> | **0.992** |
| ACAN [22] | - | <u>0.138</u> | 0.518 | <u>0.826</u> | 0.960 | 0.989 |
| SMDUN(Ours) | **0.058** | **0.137** | 0.585 | 0.820 | 0.960 | 0.989 |

### 4.1 深度估计数据集

目前，深度理解与估计网络大多在以下两个数据集上进行实验，分别是 NYUv2 数据集[30]和 KITTI

数据集[31]。

NYUv2 数据集提供了 464 个室内场景的 RGB-D 数据，由 Kinect 相机拍摄，包括了共 408k 个彩色图像和深度图像对，其中图像的分辨率都为 640×480。我们采用了文献[9]定义的训练集与测试集划分方法，在 NYUv2 数据集的 464 个场景中挑选出 249 个场景用于训练，其余 215 个场景用于测试。最终从训练场景中抽取 50k 对彩色图像和深度图像作为训练数据，并使用 654 对彩色图像与深度图像用于测试，并对深度图中空缺的区域进行填补，深度值的上限设定为 10 米。训练过程中，我们使用双线性降采样方法改变 NYUv2 数据集的图像分辨率为 256×352，作为 SMDUN 的输入和标签数据的默认分辨率。在测试阶段，我们将网络的预测深度图恢复回原始图像的大小，并在文献[9]定义的指定区域中计算预测结果的定量指标。

KITTI 是一个包含双目立体图像和 3D 点云数据的室外场景数据集，包括市区、乡村、高速公路、校园等 56 个不同的场景，其中图像的分辨率为 1241×376。我们采用文献[9]的训练集与测试集划分方法，从 56 个场景中挑选出 28 个场景用于训练，其余 28 个场景用于测试。再从训练场景中抽取 28k 个彩色图像和深度图像对作为训练数据，在测试场景中选择 697 对彩色图像与深度图像用于测试。我们对稀疏的深度图进行填补[9]并将深度图像的上限设定为 80 米，我们去掉深度图上层区域激光雷达扫描不到的部分，并使用双线性降采样方法改变 KITTI 数据集的图像分辨率至 256×512，作为 SMDUN 的输入和标签数据的默认分辨率。在测试阶段，我们将网络的预测深度图恢复到原始图像的大小，并在文献[9]定义的指定区域中计算预测结果的定量指标。

### 4.2 实验细节

SMDUN 网络采用 TensorFlow 深度学习框架，使用一块 NVIDIA RTX 2080Ti 进行训练与测试。SMDUN 的第一个编码器网络为 Resnet-50 并使用 ILSVRC[32]中的预训练模型进行初始化。

SMDUN 网络的训练过程可分为两步，在第一步训练侧重于 SMDUN 网络中的场景模态损失，其中公式(14)的参数 $\alpha_{im}$ 和 $\alpha_{mod}$ 分别设置为 $\alpha_{im}=10^{-4}$，$\alpha_{mod}=1$。网络的参数更新使用 Adam 优化算法，设置 Adam 算法的学习率为 $10^{-4}$，Adam 算法的参数设置为 $\beta_1=0.9$，$\beta_2=0.999$。第二步训练侧重于连续标签损失，公式(14)的参数 $\alpha_{im}$ 和 $\alpha_{mod}$ 分别设置为 $\alpha_{im}=1$，$\alpha_{mod}=10^{-2}$。Adam 优化算法的学习率在迭代中采用多项式衰减的策略，初始学习率设置为 $10^{-4}$，终止学习率设置为 $10^{-5}$，多项式衰减参数 power 设置为 0.9。在 NYUv2 数据集中，第一步的训练 epoch 为 6，第二步的训练 epoch 为 35，网络的 batch 大小设置为 6；在 KITTI 数据集中，第一步的训练 epoch 为 3，第二步的训练 epoch 为 35，网

**表 3 多种方法在 KITTI 数据集上的定量对比结果。表中最佳值以加粗表示，第二好的值以下划线突出。**
**Table 3 Quantitative comparison results of multiple methods on KITTI. The best value in the table is expressed in bold, and the second best value is highlighted underlined.**

| Method | AbsRel↓ | SqRel↓ | RMSE↓ | RMSElog↓ | $\delta_1$↑ | $\delta_2$↑ | $\delta_3$↑ |
|---|---|---|---|---|---|---|---|
| Eigen *et al.* [9] | 0.203 | 1.548 | 6.307 | 0.282 | 0.702 | 0.890 | 0.958 |
| Godard *et al.* [20] | 0.114 | 0.898 | 4.935 | 0.206 | 0.861 | 0.949 | 0.976 |
| Kuznietsov *et al.* [35] | 0.113 | 0.741 | 4.621 | 0.189 | 0.862 | 0.960 | 0.986 |
| Li *et al.* [34] | <u>0.109</u> | - | 4.687 | - | 0.856 | 0.962 | <u>0.988</u> |
| Cao *et al.* [12] | 0.115 | - | 4.712 | 0.198 | <u>0.887</u> | 0.963 | 0.982 |
| Xu *et al.* [36] | 0.122 | 0.897 | 4.677 | - | 0.818 | 0.954 | 0.985 |
| Gan *et al.* [37] | **0.098** | <u>0.666</u> | **3.933** | <u>0.173</u> | **0.890** | <u>0.964</u> | 0.985 |
| GASDA [21] | 0.149 | 1.003 | 4.995 | 0.227 | 0.824 | 0.941 | 0.973 |
| SMDUN(Ours) | **0.098** | **0.589** | <u>4.231</u> | **0.169** | 0.885 | **0.969** | **0.989** |

络的 batch 大小设置为 4。

我们将结果与其他流行的模型进行比较。我们把所提出的网络与下列深度网络进行了对比：Eigen 等人[9]、Liu 等人[11]、Cao 等人[12]、Fu 等人[13]、Zhang 等人[14]、Li 等人[17]、和 Godard 等人[20]、Zhao 等人[21]、Chen 等人[22]、Laina 等人[29]、Wang 等人[33]、Li 等人[34]、Kuznietsov 等人[35]、Xu 等人[36] 和 Gan 等人[37]。

### 4.3 定量评价指标

我们将 SMDUN 的结果与其他模型在以下 6 种定量指标上比较，它们的定义如下：

1）绝对相关误差（Absolute Relative Error，简称 Abs Rel）：$\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}$

2）均方相关误差（Squared Relative Error，简称 Sq Rel）：$\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|^2}{y_i}$

3）均方根误差（Root Mean Squared Error，简称 RMSE）：$\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|^2}$

4) 对数均方根误差（Root Mean Squared Error in log space，简称为 RMSElog）：$\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\log \hat{y}_i - \log y_i|^2}$

5)对数平均误差（Mean log10 Error，简称为 MLog10E 误差）：$\frac{1}{n}\sum_{i=1}^{n}|\log_{10} \hat{y}_i - \log_{10} y_i|$

6) 阈值准确度$\delta_1$、$\delta_2$和$\delta_3$：

$$\begin{cases}\delta_1: 满足 \max\left\{\frac{y_i}{\hat{y}_i},\frac{\hat{y}_i}{y_i}\right\} < 1.25 的 \hat{y}_i 数量占 n 的比例\\ \delta_2: 满足 \max\left\{\frac{y_i}{\hat{y}_i},\frac{\hat{y}_i}{y_i}\right\} < 1.25^2 的 \hat{y}_i 数量占 n 的比例\\ \delta_3: 满足 \max\left\{\frac{y_i}{\hat{y}_i},\frac{\hat{y}_i}{y_i}\right\} < 1.25^3 的 \hat{y}_i 数量占 n 的比例\end{cases}$$

其中$y_i$是标签图像中的深度值，$\hat{y}_i$为网络预测的深度图中的值，$n$为图像的像素个数。

### 4.4 实验分析

#### 4.4.1 网络剥离实验

为验证本文提出的多个模块分别具备提升网络的深度理解性能的能力，我们分别对模块进行了剥离形成了 3 种不同的网络结构，然后在 NYUv2 数据集上分别进行了定量实验和标准的 SMDUN 进行对比（见表 1）。Type-1 是只保留堆叠沙漏结构搭建的网络，其中我们整体剥离了所有的 SMFEM 子模块；Type-2 是在堆叠沙漏结构搭建的总网络上只保留图 1 中的 SMFEM-3 模块形成的网络；Type-3 是保留堆叠沙漏网络和全部 SMFEM 模块，但去除了 SMFEM 中的 ECM 部分和局部极大似然译码部分的网络；表 1 中最后一行是标准的 SMDUN 网络，保留了所有子模块和部分。

剥离实验的结果如表 1 所示。可以发现，从 Type-1 到完整的 SMDUN，随着并入的场景模态层数的增加、以及网络子模块的逐渐增加，深度理解的性能逐步提升。完整的 SMDUN 在所有指标上都得到了最优的深度估计性能，证实了所提出的 SMFEM 模块、ECM 子模块、MLDOM 子模块的有效性。

#### 4.4.2 与其他深度估计网络的比较

1）NYUv2 数据集结果

表 2 显示了 SMDUN 网络和其他先进方法在 NYUv2 数据集上的定量实验对比，表 2 中其他方法的结果是从他们原本的论文中直接摘录的。在 6 个定量深度图指标上，本文的 SMDUN 属于性能最好的第一梯队网络，在最重要的 Mlog10E 和 AbsRel 这两项指标中达到了第一，展示了其有效性。

在 NYUv2 数据集上的一些深度预测定性对比结果如图 6 所示。Eigen 等人的方法（第 3 列）能够得到三维空间的大致结构但误差较大，存在物体边缘模糊的问题；Laina 等人的方法得到的深度图像总体误差相对较低，但存在深度信息过于平滑，场景中的较小物体难以分辨并且场景中物体的轮廓存在不合理的形变的问题；DORN 得到的深度图像整体模糊、丢失了不少细节并且存在明显的网格效应；我们的方法 SMDUN 与真实图像更接近，得到的深度图像包含更多的细节信息，场景中的物体具有更清晰的轮廓。

2）KITTI 数据集结果

表 3 列出了 SMDUN 网络和其他先进方法在 KITTI 数据集上的定量对比结果，表 3 中其他方法的结果是从原本的论文中直接摘录的。为确保公平的对比条件，我们均使用稀疏的深度图作为测试数

据，深度值上限均设置为 80 m。本文的 SMDUN 在 7 个定量指标中的 5 个取得最优的结果，1 个取得次优。这表明我们的网络是解决单目 RGB 图像的深度理解问题的一个有效方法。

在 KITTI 数据集上的定性对比结果如图 7 所示，Godard 等人的方法得到的结果具有较清晰的物体轮廓，但其与真实深度图像标签存在较大的误差；DORN 的定性结果整体模糊并存在明显的网格效应；我们的方法 SMDUN 与真实深度图像在定性结果上更接近，所估计的深度图像包含更多的细节信息，场景中的物体具有更清晰的轮廓。

## 5 结束语

本文提出了一种新的单目深度理解模型 SMDUN。它以堆叠沙漏网络为总框架，同时使用不同分辨率的场景模态离散标签用于指导网络每一层级的特征。为减少由于网络加深而造成的误差积累，我们引入了有序回归码和极大似然译码的思想，进一步优化了对离散标签的学习过程。通过在堆叠沙漏网络中多次插入设计的 SMFEM 子模块，实现了逐次使用场景模态特征提取并最终基于一个综合损失函数指导网络从低层级到高层级理解深度信息。在 SMFEM 中我们设计了误差纠正子模块和极大似然译码优化子模块用于修正中间层级的错误特征，减少误差累计。在同样测试环境和评价指标下，本文方法在 NYUv2 数据集上的 AbsRel 误差最小，比参与比较的八种方法中的第二名 ACAN[22]降低了 0.72%，本文方法的 Mlog10E 误差也达到了最小，比第二名 Cao 等人[12]降低了 3.33%。本文方法在 KITTI 数据集上的 SqRel 误差最小，比参与比较的八种方法中的第二名 Gan 等人 [37]降低了 11.56%，RMSElog 误差则比 Gan 等人 [37]降低了 2.31%。实验表明本网络在单目深度估计任务中，可以提取到更精确的深度信息。从定性结果而言，本文方法在预测出的深度图像中相比[9]，DORN[13,20,29]，包含更多的图像细节，并且保持了较好的目标边缘特性。

未来研究中，我们将进一步把极大似然译码优化模块广义化到其他深度学习的任务中，争取使本文方法不仅能够适用于解决深度理解问题，还能够对解决语义分割、人体关节点检测等问题有所帮助。

**参考文献**


[1] Fang Yajun, Masaki Ichiro, Horn Berthold. Depth-based target segmentation for intelligent vehicles: Fusion of radar and binocular stereo[J]. IEEE Transactions on Intelligent Transportation Systems, 2002, 3(3): 196-202.

[2] Kao Jiun-Yu, Tian Dong, Mansour Hassan, Vetro Anthony, Ortega Antonio. Moving object segmentation using depth and optical flow in car driving sequences[C]//2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 11-15.

[3] Biswas Joydeep, Veloso Manuela. Depth camera based indoor mobile robot localization and navigation[C]//2012 IEEE International Conference on Robotics and Automation. IEEE, 2012: 1697-1702.

[4] Cui Jiyun, Zhang Hao, Han Hu, Shan Shiguang, Chen Xilin. Improving 2D face recognition via discriminative face depth estimation[C]//2018 International Conference on Biometrics (ICB). IEEE, 2018: 140-147.

[5] Saxena Ashutosh, Chung Sung H, Ng Andrew Y. Learning depth from single monocular images[C]//Advances in Neural Information Pocessing Systems. 2006: 1161-1168.

[6] Zhang Ruo, Tsai Ping-Sing, Cryer James Edwin, Shah Mubarak. Shape-from-shading: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(8): 690-706.

[7] Nayar S K, Nakagawa Y. Shape from focus[J]. IEEE Transactions on Pattern analysis and Machine Intelligence, 1994, 16(8): 824-831.

[8] Favaro P, Soatto S. A geometric approach to shape from defocus[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 406-417.

[9] Eigen David, Puhrsch Christian, Fergus Rob. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in Neural Information Pocessing Systems. 2014: 2366-2374.

[10] Xie Junyuan, Girshick Ross, Farhadi Ali. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks[C]//European Conference on Computer Vision. Springer, Cham, 2016: 842-857.

[11] Liu Fayao, Shen Chunhua, Lin Guosheng, Reid Ian. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE Transactions on

Pattern Analysis and Machine Intelligence, 2015, 38(10): 2024-2039.

[12] Cao Yuanzhouhan, Wu Zifeng, Shen Chunhua. 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.

[13] Fu Huan, Gong Mingming, Wang Chaohui, Batmanghelich Kayhan, Tao Dacheng. Deep ordinal regression network for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2002-2011.

[14] Zhang Zhenyu, Xu Chunyan, Yang Jian, Gao Junbin, Cui Zhen. Progressive hard-mining network for monocular depth estimation[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3691-3702.

[15] Newell Alejandro, Yang Kaiyu, Deng Jia. Stacked hourglass networks for human pose estimation[C]. European Conference on Computer Vision, 2016: 483-499.

[16] Q. C, Cover T M, Thomas Joy A . Elements of information theory[J]. Publications of the American Statal Association, 2006, 103(481):429-429.

[17] Li Jun, Klein Reinhard, Yao Angela. A two-streamed network for estimating fine-scaled depth maps from single rgb images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3372-3380.

[18] Xu Dan, Ricci Elisa, Ouyang Wanli, Wang Xiaogang, Sebe Nicu. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5354-5362.

[19] Garg Ravi, BG Vijay Kumar, Carneiro Gustavo, Reid Ian. Unsupervised CNN for single view depth estimation: Geometry to the rescue[C]. European Conference on Computer Vision, 2016: 740-756.

[20] Godard Clément, Mac Aodha Oisin, Brostow Gabriel J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279.

[21] Zhao Shanshan, Fu Huan, Gong Mingming, Tao Dacheng. Geometry-aware symmetric domain adaptation for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9788-9798.

[22] Chen Yuru, Zhao Haitao, Hu Zhengwei. Attention-based context aggregation network for monocular depth estimation[J]. arXiv preprint arXiv:1901.10137, 2019.

[23] Islam Md Amirul, Naha Shujon, Rochan Mrigank, Bruce Neil, Wang Yang. Label refinement network for coarse-to-fine semantic segmentation[J]. arXiv preprint arXiv:1703.00551, 2017.

[24] Zhang Feihu, Prisacariu Victor, Yang Ruigang, Torr Philip H.S. Ga-net: Guided aggregation net for end-to-end stereo matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 185-194.

[25] Eigen David, Fergus Rob. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 2650-2658.

[26] Xie Saining S, Tu Zhuowen. Holistically-Nested Edge Detection[J]. International Journal of Computer Vision, 2017, 125(5):3-18.

[27] Lin Guosheng, Milan Anton, Shen Chunhua, Reid Ian. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1925-1934.

[28] Li Dawei, Wang Sifan, Tang Xue-song, Kong Weijian, Chen Yang. Double-stream atrous network for shadow detection[J]. Neurocomputing, 2020, 417: 167-175.

[29] Laina Iro, Rupprecht Christian, Belagiannis Vasileios, Tombari Federico, Navab Nassir. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016: 239-248.

[30] Silberman Nathan, Hoiem Derek, Kohli Pushmeet, Fergus Rob. Indoor segmentation and support inference from rgbd images.[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012:746-760.

[31] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics:The KITTI dataset[J]. The International Journal

of Robotics Research, 2013, 32(11):1231-1237.

[32] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, Berg Alexander C, Fei-Fei Li. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[33] Wang Peng, Shen Xiaohui, Lin Zhe, Cohen Scott, Price Brian, Yuille Alan. Towards unified depth and semantic prediction from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. 2015: 2800-2809.

[34] Li Bo, Dai Yuchao, He Mingyi. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference[J]. Pattern Recognition, 2018, 83: 328-339.

[35] Kuznietsov Yevhen, Stuckler Jrg, Leibe Bastian. Semi-supervised deep learning for monocular depth map prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6647-6655.

[36] Xu Dan, Wang Weri, Tang Hao, Liu Hong, Sebe Nicu. Structured attention guided convolutional neural fields for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3917-3925.

[37] Gan Yukang, Xu Xiangyu, Sun Wenxiu, Lin Liang. Monocular depth estimation with affinity, vertical pooling, and label enhancement[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 224-239.