计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330485
61(5): 1336−1347, 2024

联合语义分割的自监督单目深度估计方法

宋霄罡[1,2] 胡浩越[1] 宁靖宇[1] 梁 莉[1] 鲁晓锋[1,2] 黑新宏[1,2]
[1](西安理工大学计算机科学与工程学院 西安 710048)
[2](人机共融智能机器人陕西省高校工程研究中心(西安理工大学) 西安 710048)
(songxg@xaut.edu.cn)

Self-Supervised Monocular Depth Estimation Method for Joint Semantic Segmentation

Song Xiaogang[1,2], Hu Haoyue[1], Ning Jingyu[1], Liang Li[1], Lu Xiaofeng[1,2], and Hei Xinhong[1,2]
[1](*School of Computer Science and Engineering*, *Xi'an University of Technology*, *Xi'an* 710048)
[2](*Human Machine Integration Intelligent Robot Shaanxi Provincial University Engineering Research Center* (*Xi'an University of Technology*), *Xi'an* 710048)

Abstract In this paper, the mutually beneficial relationship between depth estimation and semantic segmentation is investigated, and a self-supervised monocular depth estimation method for joint semantic segmentation USegDepth is proposed. The shared encoder for semantic segmentation and depth estimation is implemented to achieve semantic guidance. To further improve the across multiple tasks performance of the encoder, a multi-task feature extraction module is designed. The module is stacked to generate the shared encoder, solving the poor feature representation problem of the model due to limited receptive field and lack of cross-channel interaction, and the model accuracy is improved further. And a cross-task interaction module is proposed for bidirectional cross-domain information interaction to refine the depth features, improving depth estimation performance, especially in weak texture regions and object boundaries with limited luminosity consistency supervision. Through training and evaluation on KITTI dataset, the experimental results show that the mean square relative error of USegDepth is reduced by 0.176 percentage points compared with that of SGDepth, and the threshold accuracy reaches 98.4% at a threshold value of 1.25^3, proving the high accuracy of USegDepth in depth prediction.

Key words self-supervised deep learning; depth estimation; semantic segmentation; multi-task association; shared encoder

摘 要 研究深度估计和语义分割的图像之间的互利关系,提出了一种联合语义分割的自监督单目深度估计方法 USegDepth. 语义分割和深度估计任务通过共享编码器,实现语义引导. 为了进一步提高编码器的跨多任务性能,设计了多任务特征提取模块,堆叠该模块构成共享编码器,解决有限感受野和缺乏跨通道交互导致的模型特征表示能力欠佳问题,进一步提升模型精度. 同时,提出跨任务交互模块,通过双向的跨域信息交互细化特征表示,提升深度估计表现,特别是光度一致性监督有限的弱纹理区域和物体边界. 通过在 KITTI 数据集上的训练和全面评估,实验结果显示所提的 USegDepth 模型方法的均方相对误差相比于 SGDepth 降低了 0.176 个百分点,在阈值为 1.25^3 的阈值精度达到了 98.4%,证明了 USegDepth 在深度预测上具有较高的准确率.

收稿日期: 2023-06-12; **修回日期**: 2024-01-03
基金项目: 国家自然科学基金面上项目 (52372418, 62076201); 国家自然科学基金联合基金项目 (U1934222); 陕西省重点研发计划项目(2022GY-245)
This work was supported by the General Program of the National Natural Science Foundation of China (52372418, 62076201), the National Natural Science Foundation Joint Fund Project (U1934222), and the Key Research and Development Plan of Shaanxi Province (2022GY-245).

关键词 自监督深度学习；深度估计；语义分割；多任务联合；共享编码器

中图法分类号 TP39

深度估计和语义分割作为场景理解的2个重要组成部分，在计算机视觉领域受到了广泛的关注. 近年来，深度卷积神经网络凭借其优异的深度特征表示能力，极大地提高了深度估计和语义分割任务的性能.

深度估计是计算机视觉中的一个热门研究方向，在3维重建、场景理解、环境感知等任务中起到了关键作用[1]. 许多方法利用不同类型的传感器获得深度信息，包括单目相机[2]、多视角相机[3]和激光雷达[4]. 虽然单目相机价格低廉且应用广泛，但由于其尺度模糊性，也是最具挑战性的. 目前表现较好的单目深度估计方法[5]都依赖于有监督训练，需要大量的深度真值标签用于建立监督信号，而获取深度真值标签成本高昂，并且存在明显噪声. 为了避免这些缺点，一些自监督方法利用无标签立体图像[6]或单目视频帧[7]作为输入，基于光度一致性原理，将光度一致性损失作为监督信号，取得了与监督方法相当的性能. 然而，大部分的自监督深度估计方法主要依赖光度损失和平滑损失作为监督信号，因此，在低纹理区域、反射表面和存在遮挡的区域存在高模糊性，得到了较差的深度估计表现. 为了减少这些病态区域的影响，许多工作都尝试引入来自其他领域的约束，如光流[8]、曲面法线[9]以及语义信息[10]. 光流通过区分相机运动和物体运动，处理深度估计中违反了光度一致性假设的移动物体；而曲面法线可以矫正深度图在平面区域发生的畸变，利用几何关系进行更精确的估计；语义信息则是从语义实例中过滤动态对象或显式地建模对象运动，将其合并到场景几何图形中或者通过正则化相应语义对象内的深度平滑度，加强深度和语义预测之间的一致性.

本文设计了一个跨任务交互模块(cross tasks interactive module, CTIModule). 通过计算参考特征和目标特征之间的相似性，有效地利用异构表示之间的跨任务交互作用，将深度特征和语义信息之间的关系量化为一种注意形式，细化与语义一致的深度特征，从而产生更好的深度预测. 在特征提取阶段，为了实现语义引导，深度估计和语义分割任务共享同一个编码器，同时，利用提出的多任务特征提取模块(multi-task feature extraction module, MTFEModule)堆叠构建共享编码器，进一步改善因感受野大小和缺乏跨通道交互导致的特征表示能力下降，提升编码器的特征提取能力.

本文的主要贡献有3个方面：

1)提出了一种新的联合语义分割任务的自监督单目深度估计方法，其中深度估计任务与语义分割任务共享编码器，为了进一步提高共享编码器的特征表示能力，提出多任务特征提取模块，利用多路径和注意力融合策略，使神经元自适应地根据输入选择感受野大小，提升深度估计表现；

2)提出跨任务交互模块，充分学习语义的隐式表示，细化与语义一致的深度特征，产生具有语义感知的深度信息，相应地，利用深度特征表示中的场景几何信息进行语义特征增强，最终实现了深度估计和语义分割任务的共同提升；

3)在KITTI数据集上，基于公用评判标准，进行了全面的性能评估，证明了本文提出的USegDepth模型在自监督单目深度估计中的优秀表现，并利用相应的消融实验证明了所提策略的有效性.

1 相关工作

传统深度估计方法中大多采用多视图图像作为输入，包括使用双目或多目相机获取的图像对以及使用单目相机获取的连续图像帧. 多视角立体视觉和运动恢复结构是2种经典的传统多视图深度估计方法. 但基于传统方法的单目深度估计大都依赖场景中特殊视觉信息，因此对环境要求严格. 并且，所需计算资源大，难以满足实际工业需求.

随着深度学习的发展，很大程度上弥补了传统算法存在的局限性. 深度学习策略可以分为有监督深度估计方法和无监督深度估计方法.

有监督深度学习是指使用传感器获得的深度真值建立网络训练的监督信号. Eigen等人[11]利用卷积神经网络实现单目深度估计，提出一个双尺度网络结构，其中细尺度网络将粗尺度网络的全局预测结果与原始输入图片的局部信息结合，丰富局部细节，使场景中的边缘信息更加突出. 相较于传统方法，基于有监督学习的深度估计方法不依赖图像对或连续图像序列之间的特征点提取与匹配，降低了场景限制，并且实现了高质量的深度估计结果. 但是，基于有监督学习的网络模型训练依赖深度真值标签建立监督信号，而深度值标签的获取成本高昂且不易采

集，导致其在实际使用中受限.

无监督深度估计方法在训练模型的过程中无需通过激光雷达等设备获取到真实深度建立监督信号，仅依赖 RGB 图像学习，极大降低了深度估计任务成本，并且获得了可以与有监督方法相媲美的估计精度. Zhou 等人[12]提出了一个基于单目视频流的深度估计框架. 利用深度估计网络预测的深度图和位姿估计网络计算的相对位姿，进行视图重建. 理想状态下，重建视图应与原视图生成一致的光度信息，利用二者之间的光度差异建立光度一致性损失，作为模型训练的主要监督信号.

目前单目深度估计和语义分割的研究都取得了一定成果，但大多数方法都在强调学习鲁棒的回归，忽略了二者之间的相互作用. 这 2 个任务之间具有共同的特性，例如，语义分割和深度估计都可以揭示场景布局、对象的形状以及边缘. 联合语义相关任务进行深度估计的方法可以分为 3 类：语义和深度分支共享权重、混合语义和深度特征以及建模语义与深度之间的统计学关系. Wang 等人[13]提出一种联合语义分割训练的深度估计方法，包括一个预测稠密深度图的全局网络和一个计算语义标签的区域估计网络，经过 2 层条件随机场优化得到细化后的输出. Jafari 等人[14]通过量化深度预测图和语义预测图之间的跨膜态影响，提出一种由估计模块和细化模块组成的模块化神经网络. 其中估计模块根据不同的任务，分别构建相应的子网络，2 个子网络共同训练，相互促进. Liu 等人[15]利用语义分割生成的语义标签，判断像素或区域所属的语义类，引导深度估计. Zhu 等人[16]提出利用语义分割的边缘指导深度边缘，有效改善了物体边缘质量. 张玉亮等人[17]构建以超像素为节点的拓扑关系图，有效融合边缘语义信息，提高了提取局部结构信息的能力.

2 USegDepth 模型

2.1 整体框架

图 1 描述了本文提出的 USegDepth 模型的整体框架. 在整体框架中，一共有 3 个子任务，包括深度估计、位姿估计以及语义分割，这 3 个子任务共同训练. 其中位姿估计任务有独立的编码器和解码器，深度估计任务与语义分割任务共享一个编码器，但有各自独立的解码器. 共享编码器利用提出的多任务特征提取模块构建，跨任务交互模块嵌入于深度估计和语义分割的解码器各层之间，实现 2 个特定任务解码器之间的跨任务交互. 框架的输入为连续视频帧 I_{t-1} 以及 I_t，深度估计与语义分割网络每次以 1 张视频帧作为输入，经过共享编码器提取特征后，将提取的特征分别送入各自独立的解码器中，深度估计的解码器最终输出 1 张稠密深度图，语义分割的解码器则生成 1 张相应的语义分割图. 位姿估计网络将 2 张相邻帧聚合后作为输入，计算 2 帧之间的相对位姿，利用计算得到的深度图和位姿进行视图重建，根据光度一致性原理，建立重建视图与目标视图之间的光度损失作为主要的监督信号. 同时，利用语义标签与模型预测生成的语义分割图建立交叉熵损失，通过反向传播提供更多的语义感知表示，优化深度

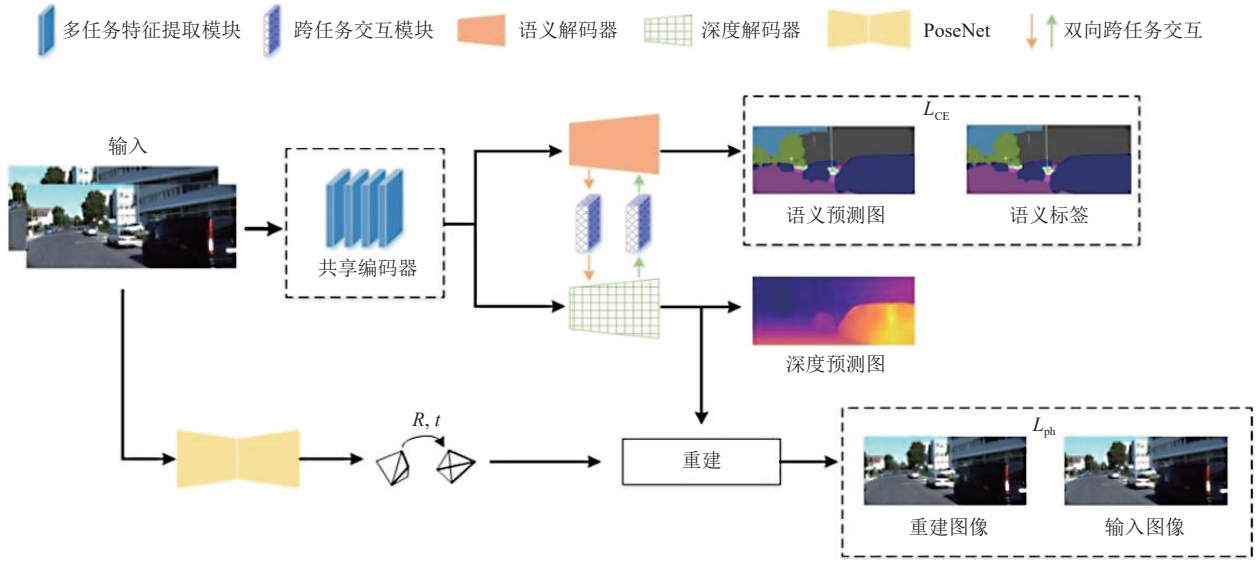


Fig. 1 Overall framework of USegDepth

图 1 USegDepth 整体框架

估计网络层. 并且我们还联合了语义引导的三重态损失，利用场景语义的局部几何关系，优化深度图边界，使输出的深度图与语义边界更加对齐.

2.2 多任务特征提取模块

为了构建一个具有改进特征表示能力的共享编码器，提高编码器跨多任务性能，本文提出了多任务特征提取模块，如图 2 所示. 其中多任务特征提取模块包括分组特征映射和多分支特征注意力融合 2 个操作. 首先将每个块的特征映射沿着通道维度划分为多个组和更细粒度的子组，每个子组的分支上采用不同尺度的卷积核提取特征，然后利用注意力融合模块对具有不同尺度卷积核的分支生成的多尺度特征分配不同的权重，其中每个组的特征表示是其子组表示的加权组合. 通过堆叠多任务特征提取模块，构建了深度估计与语义分割的共享编码器.

图 2(a)描述了一个多任务特征提取模块的整体结构，图 2(b)是其中注意力融合模块的具体结构表示. 分组特征操作，采取了多分支特征提取的策略. 将特征分为 K 个组，同时，在一个特征映射组中将特征再次分组，引入了一个新的超参数 r，特征组的总数 $G=Kr$. 本文设置 K=2, r=2. 之后是对不同分支的特征进行注意力特征融合. 结合局部特征和全局特征，在注意力模块中对多个分支实现通道交互. 通过对 2 个不同卷积核大小的网络分支分配不同的注意力，使神经元可以自适应地根据不同输入信息调整其感受野大小. 为了尽可能保持注意力融合模块的轻量级，只是将局部上下文添加到注意力模块内的全局上下文中. 全局通道上下文表示如式(1)所示：

$$P(x)=\beta(PWConv_2(\delta(\beta(PWConv_1(g(x)))))),\quad (1)$$

$$g(x)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_{[:,i,j]}.$$

其中 x 表示输入的中间特征，$g(x)$ 表示全局平均池化操作，β 表示批处理正则化，δ 表示 ReLU 激活函数. 局部通道上下文聚合器选择点向卷积(point-wise convolution, PWConv)实现每个空间位置的点向通道交互. 为了节约参数，通过瓶颈结构计算局部通道上下文 $\boldsymbol{L}(x)\in\mathbb{R}^{C\times H\times W}$，如式(2)所示：

$$\boldsymbol{L}(x)=\beta(PWConv_2(\delta(\beta(PWConv_1(x))))).\quad (2)$$

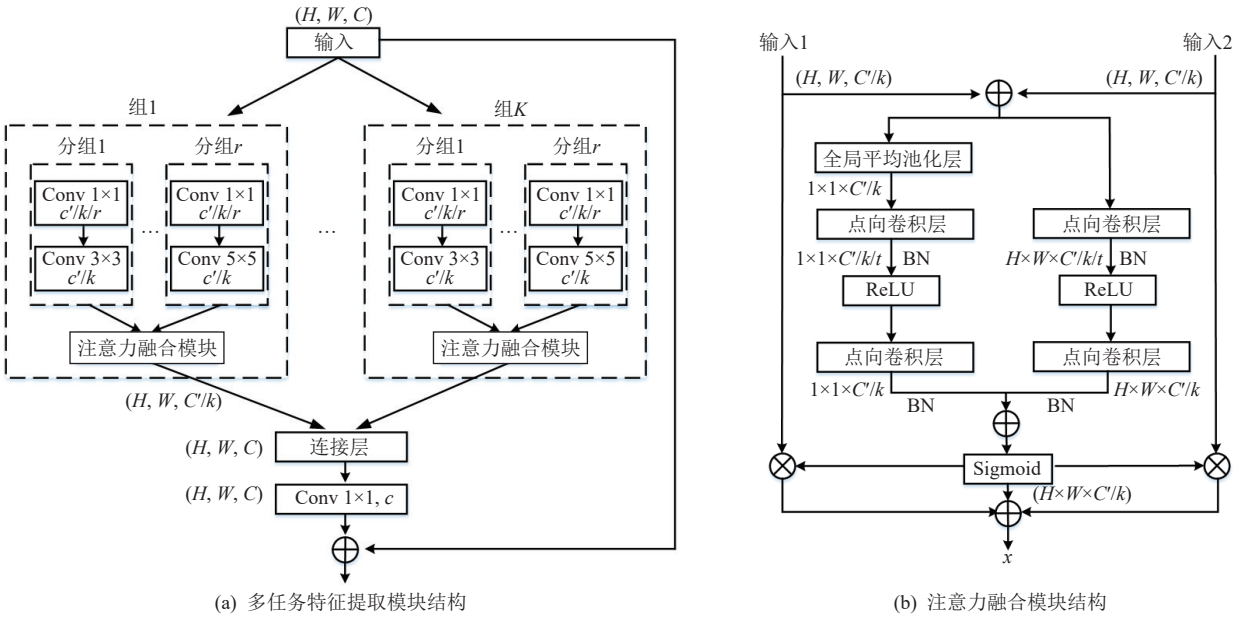
$\boldsymbol{L}(x)$ 与输入特征具有相同的形状，它可以保留和突出低级特征中的细节信息. 将全局通道上下文 $g(x)$ 和局部通道上下文 $\boldsymbol{L}(x)$ 聚合后，最终得到细化后的特征 x'，如式(3)所示：

$$x'=x\otimes\boldsymbol{M}(x)=x\otimes\sigma(\boldsymbol{L}(x)\oplus g(x)),\quad (3)$$

其中，$\boldsymbol{M}(x)\in\mathbb{R}^{C\times H\times W}$ 为注意力特征融合模块产生的注意力权重，σ 表示 Sigmoid 函数，$\oplus$ 代表广播加法，$\otimes$ 为元素级乘法.

2.3 跨任务交互模块

跨任务交互模块的结构如图 3 所示，该模块位于深度估计解码器和语义分割解码器的每个解码器层之间，利用来自 2 个解码器之间的特征进行交互信息，产生具有语义感知的深度特征，细化深度预测，



(a) 多任务特征提取模块结构　　(b) 注意力融合模块结构

Fig. 2 Multi-tasks feature extraction module

图 2 多任务特征提取模块

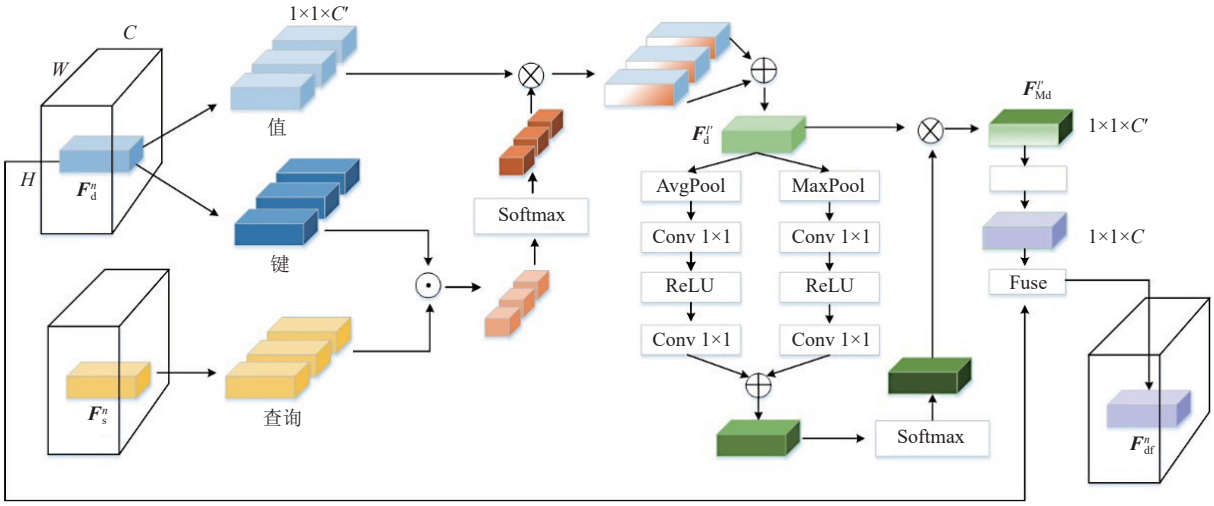


Fig. 3 Cross-task interaction module structure

图 3 跨任务交互模块结构

同时, 也可以利用深度特征中的场景几何信息增强语义特征. 每个跨任务交互模块有 2 个输入, 分别为深度估计的特征图和语义分割的特征图, 固定其中一个为参考特征, 另一个为目标特征, 利用参考特征细化目标特征. 由于跨任务交互模块是单向的数据流, 因此, 为了实现双向特征增强, 每个解码器层之间都同时嵌入 2 个跨任务交互模块. 以下描述一个跨任务交互模块的具体实现: 固定深度估计编码器生成的特征为目标特征、语义分割编码器生成的特征为参考特征. USegDepth 框架中, 深度解码器与语义分割解码器都由 n 个解码层组成(其中 n=0, 1, 2, 3, 4), 每层解码器中都包含上采样操作, 特征图的空间分辨率会逐层增加 1 倍, 即第 n 层解码器的深度特征图 $\boldsymbol{F}_{\mathrm{d}}^n$ 与语义分割特征图 $\boldsymbol{F}_{\mathrm{s}}^n$ 的空间分辨率为 $H/2^{4-n}$, $W/2^{4-n}$, 其中 H 和 W 为输入图像的高度和宽度.

跨任务交互模块附加在深度估计和语义分割的每个解码器层之间, 首先通过计算 2 个特征映射 $\boldsymbol{F}_{\mathrm{d}}^n$ 和 $\boldsymbol{F}_{\mathrm{s}}^n$ 之间的跨任务相似度, 作为像素级的注意力得分, 使深度特征感知语义信息. 由于任务性质的不同, 使用原始特征向量直接计算相似度是不可行的, 因此, 采用了一个线性投影 ρ, 将原始特征从输入维度 C 变换为 C'. 应用 3 个独立的线性嵌入, 每个嵌入分别作为查询 ρ_{q}、键 ρ_{k} 和值 ρ_{v} 函数. 目标特征映射 $\boldsymbol{F}_{\mathrm{d}}^l$ 作为键和值嵌入的输入, 参考特征映射 $\boldsymbol{F}_{\mathrm{s}}^l$ 作为查询嵌入的输入. 通过表示子空间间接计算跨任务相似度, 细化后的特征的计算方法如式(4)所示:

$$\boldsymbol{F}_{\mathrm{d}}^n(i)' = \beta(A(i)) \times \rho_{\mathrm{v}}\left(\boldsymbol{F}_{\mathrm{d}}^n(i)\right), \tag{4}$$

$$A(i) = \frac{\rho_{\mathrm{k}}\left(\boldsymbol{F}_{\mathrm{d}}^n(i)\right)^{\mathrm{T}} \rho_{\mathrm{q}}\left(\boldsymbol{F}_{\mathrm{s}}^n(i)\right)}{\sqrt{C'}},$$

其中归一化因子 β 用于缩放输入, 注意力分数 $A(i)$ 由相似度量化得到.

分配与语义具有高度一致的深度特征较大的注意力分数, 利用语义信息细化深度特征. 解码器中放置了 2 个跨任务交互模块, 当语义特征作为目标特征时, 则深度特征成为参考特征, 利用深度特征映射细化语义分割的特征表示.

受多头注意的启发, 采用多重线性投影, 通过不同的表示子空间计算特征向量之间的相似性, 更有效地从多嵌入中选择性地提取深度特征, 最大限度地利用跨空间性. 具体地, 使用 H 个不同的投影函数, $\rho^h(h=1, 2, \cdots, H)$, 将查询、键和值映射到 H 个独立的子空间. 式(4)中的跨任务相似性直接扩展到多嵌入方案, 如式(5)所示:

$$\boldsymbol{F}_{\mathrm{d}}^{n^h}(i)' = \beta\left(A^h(i)\right) \times \rho_{\mathrm{v}}^h\left(\boldsymbol{F}_{\mathrm{d}}^n(i)\right), \tag{5}$$

其中

$$\beta\left(A^h(i)\right) = \frac{\mathrm{e}^{A^{h(i)}}}{\sum\limits_{h' \in H} \mathrm{e}^{A^{h'(i)}}},$$

$$A^h(i) = \frac{\rho_{\mathrm{k}}^h\left(\boldsymbol{F}_{\mathrm{d}}^n(i)\right)^{\mathrm{T}} \rho_{\mathrm{q}}^h\left(\boldsymbol{F}_{\mathrm{s}}^n(i)\right)}{\sqrt{C'}}.$$

细化的总特征映射 $\boldsymbol{F}_{\mathrm{d}}^l(i)'$ 是对每个嵌入函数细化后的特征映射的总和, 如式(6)所示:

$$\boldsymbol{F}_{\mathrm{d}}^l(i)' = \sum_{h=1}^{H} \boldsymbol{F}_{\mathrm{d}}^{n^h}(i), \tag{6}$$

其中, h 表示多重线性嵌入的索引. 采用 Softmax 作为归一化函数, 计算每个嵌入的重要性, 然后有选择地

利用来自多重注意力的输出.

随后对细化后的总深度特征映射$\boldsymbol{F}_{\mathrm{d}}^{l'}$实施通道注意力，分别利用全局最大池化和全局平均池化压缩空间维度，保留通道信息，然后再将多层感知机处理后的特征相加，利用 Sigmoid 激活函数得到最终的通道注意力权重，对$\boldsymbol{F}_{\mathrm{d}}^{l'}$再次加权，有选择地关注更有用的通道信息. 具体如式(7)所示：

$$\begin{aligned}\boldsymbol{F}_{\mathrm{Md}}^{l'}=&M_{\mathrm{c}}\left(\boldsymbol{F}_{\mathrm{d}}^{l'}\right)=\\&\beta\left(MLP\left(AvgPool\left(\boldsymbol{F}_{\mathrm{d}}^{l'}\right)\right)+MLP\left(MaxPool\left(\boldsymbol{F}_{\mathrm{d}}^{l'}\right)\right)\right).\end{aligned}\tag{7}$$

最后将输入特征图$\boldsymbol{F}_{\mathrm{d}}^{n}$与加权后的$\boldsymbol{F}_{\mathrm{Md}}^{l'}$进行特征融合. 具体地，先将细化后的特征图$\boldsymbol{F}_{\mathrm{d}}^{l'}$投影到原始维数 C，投影操作用γ表示，然后将 2 个特征映射拼接为$\left[\gamma\left(\boldsymbol{F}_{\mathrm{d}}^{l'}\right),\boldsymbol{F}_{\mathrm{d}}^{n}\right]$，通过 2 个卷积层作用于连接的特征映射，完成与初始特征图$\boldsymbol{F}_{\mathrm{d}}^{n}$的融合，从而生成$\boldsymbol{F}_{\mathrm{df}}^{n}$，$\boldsymbol{F}_{\mathrm{df}}^{n}$作为第 n+1 层深度解码器的输入.

2.4 网络结构

深度估计任务和语义分割任务网络采用编码器-解码器架构，这 2 个任务的编码器共享，采用提出的多功能特征提取模块堆叠构成，遵循残差结构实现跳跃连接，将上层网络信息引入到下层网络，解决深层网络中梯度消失问题.

解码器采用上采样层和卷积层结合的策略构建，一共包括 4 个相同的 Upconv 结构，分别为 Upconv 1, Upconv 2, Upconv 3 和 Upconv 4，每个 Upconv 中包括特征图融合、多个卷积层和上采样操作. 每个 Upconv 的输入包括上一层网络的输出和具有相同大小的编码器生成的特征图，将二者融合后进行卷积和上采样操作，最后 Upconv4 的输出图像尺寸与输入图像尺寸相同. 在解码器中，对输入矩阵扩充时采用反射填充，扩充值使用附近像素值，相较于零填充，可以减少物体边界处的模糊问题，提高特征图清晰度.

整体采用 U-Net 架构，在编码器中通过下采样收缩特征分辨率，解码器中对称地通过上采样实现特征分辨率的扩展，并结合下采样各层信息对细节信息还原，更好地提高输出图像精度.

跨任务交互模块嵌入在深度估计和语义分割网络的第 l 层解码器之间，其中 l = 0, 1, 2. 采用 H = 4 的多嵌入方案，原始特征与嵌入特征之间的维度比为 2，即$C'=2C$.

位姿网络中，同样采用编码器-解码器结构，其中编码器采用 ResNet18，因为位姿网络的输入为 2 张 RGB 图像在通道维度的拼接，因此，扩展第 1 个卷积核的输入维度为 6 通道，解码器中首先使用压缩操作对图像特征进行降维，之后进行多次卷积操作，最后预测出相机位置变化的平移运动和旋转运动.

2.5 损失函数

2.5.1 光度一致性损失

光度损失是自监督深度估计的主要监督信号. 利用深度估计网络预测的像素级深度图，以及位姿估计网络计算的 6-Dof 相对位姿，基于重建原理，实现从输入图像I_t到I_{t-1}的投影，获得投影后的像素坐标：

$$\hat{I}_t=W_t(I_{t-1},p'),\tag{8}$$

$$p'=KT_{t\to t-1}Z_tK^{-1}p,$$

其中 p 为 I_t 中像素的齐次坐标，p'为 p 经过$T_{t\to t-1}$的变换坐标. $W_t(\cdot)$是一个可微双线性采样器[18]，在I_{t-1}中获得p'附近的像素，然后在$\hat{I}_t$中找到 p 的线性插值像素. 理想情况下，深度估计网络和位姿估计网络都得到了最佳训练，I_t和$\hat{I}_t$应该是对齐的. 利用结构相似性指数度量[19]与 L_1 损失结合，建立光度一致性损失[3] L_{ph}：

$$L_{\mathrm{ph}}=(1-\alpha)\left|I_t-\hat{I}_t\right|+\alpha\frac{1-SSIM\left(I_t,\hat{I}_t\right)}{2}.\tag{9}$$

2.5.2 平滑损失

添加了边缘感知平滑损失[20] L_{sm}：

$$L_{\mathrm{sm}}=|\partial_x d_t|\mathrm{e}^{-|\partial_x I_t|}+\left|\partial_y d_t\right|\mathrm{e}^{-\left|\partial_y I_t\right|}.\tag{10}$$

2.5.3 交叉熵损失

由于联合了语义分割任务，因此，在网络训练中添加交叉熵损失. 通过逐个像素对比本文方法中预测的分割图与标签之间的差异得到损失值，具体如式(11)所示. 通过反向传播语义分割的交叉熵损失可以进一步优化深度估计层，提供更多的语义引导.

$$L_{\mathrm{ce}}=-\sum_{class}\left(y_{class}\log\left(p_{class}\right)\right),\tag{11}$$

其中 $class$ 代表类别数，y_{class}为热编码值，且$y_{class}\in\{0,1\}$，p_{class}为网络预测值经过 Softmax 得到的概率值.

2.5.4 三重态损失

基于场景语义中的局部几何关系，每个对象实例中的相邻像素具有相似的深度值，而跨语义边界的像素则具有较大的深度差异. 语义引导的三重态损失[21]通过在表示空间中定义和优化距离，为优化深度表示提供了有效的监督信号，使深度解码器在边界区域上产生更多的可鉴别性特征，从而使输出的深度图与语义边界更加对齐. 如式(12)所示：

$$L_{\mathrm{SGT}}=\frac{\sum_i\left[\left|p_i^+\right|,\left|p_i^-\right|>T\right]L_{p_i}}{\sum_i\left[\left|p_i^+\right|,\left|p_i^-\right|>T\right]},\tag{12}$$

$$L_{p_i} = \max\left(0, d^+(i) + m - d^-(i)\right),$$

其中p_i^+和p_i^-是局部补丁p_i中的正、负像素集，i是锚点的空间位置. 使用p_i^+和p_i^-来确定p_i是否相交于语义边界. 设置了一个阈值T, 如果$\left|p_i^+\right|$和$\left|p_i^-\right|$都大于T, 则表示p_i与物体之间的边界相交. 减少锚点与正特征之间的距离，并增加锚点与负特征之间的距离. 当区分出带有语义信息的p_i^+和p_i^-后，将位置同步到预测的深度图中，区分出d^+区域和d^-区域，$d^+(i)$和$d^-(i)$分别是2个区域关于锚点i的深度差异.

整体训练损失函数为：

$$L_{\text{total}} = L_{\text{ph}} + \beta L_{\text{sm}} + \gamma L_{\text{ce}} + \delta L_{\text{SGT}}, \tag{13}$$

其中β, γ, δ为各损失的权重参数.

3 实 验

3.1 训练细节

训练前调整原始图像分辨率为192×640, 设置批处理大小为12.使用Adam优化器，初始学习速率设置为1.5E−4, 共训练20代，经过10代和15代的训练后，学习速率衰减0.1. 遵循之前的工作[22]，设置α=0.85. 平滑损失β、交叉熵损失γ以及三重态损失δ的权重参数分别设置为：β= 0.001, γ= 0.3, δ= 0.1. 三重态损失的边缘m设置为0.3, 阈值T设为4, m和T值的设置遵循先前工作[21].

3.2 数据集

在KITTI数据集上对模型进行了深度估计测试. 使用Eigen split[22]的分割序列，包括39 910张和4 424张图像分别用于模型的训练和验证，697张图像用于模型评估. 为了训练语义分割模型，首先使用一个现有的网络[23]生成伪标签，用于评估与模型预测的语义分割图之间的差异，建立交叉熵损失. 使用与KITTI 2015[24]对应的KITTI语义分割基准训练集提供的200张RGB图像和对应的语义标签评估语义分割性能.

3.3 评价指标

基于单目深度估计研究中常用的评价指标客观评估本文所提出的模型. 评价指标主要包括错误率和准确率，以下1)~4)为错误率，5)为准确率.

1) 相对误差(absolute relative error, *AbsRel*)：

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|D_i - D_i^*\right|}{D_i^*}. \tag{14}$$

2) 均方相对误差(squared relative error, *SqRel*)：

$$SqRel = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|D_i - D_i^*\right|^2}{D_i^*}. \tag{15}$$

3) 均方根误差(root mean square error, *RMSE*)：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|D_i - D_i^*\right|^2}. \tag{16}$$

4) 对数均方根误差(root mean square logarithmic error, *RMSElog*)：

$$RMSElog = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\log D_i - \log D_i^*\right|^2}. \tag{17}$$

5) 阈值精度：

$$\delta_\tau = \max\left\{\frac{D_i}{D_i^*}, \frac{D_i^*}{D_i}\right\} < \varepsilon^\tau, \tag{18}$$

其中$\varepsilon = 1.25$, $\tau \in \{1,2,3\}$, 阈值精度所取得的阈值为1.25, 1.25^2, 1.25^3. 以上5个指标中D_i表示预测的深度值，D_i^*表示真实的深度值，N表示像素总数.

3.4 对比分析

表1展示了本文提出的USegDepth模型方法在KITTI数据集上与其他方法的深度估计对比结果，包括联合语义分割任务的方法和单任务训练方法. 对比结果显示USegDepth在每个度量上都优于其他工作. 相较于文献[30]在训练和测试过程中都需要语义信息，本文联合语义分割的多任务网络中，语义信息只用于训练阶段，在测试阶段无需额外的语义处理操作. 与USegDepth的研究类似，SGDepth[31]同样采用深度估计与语义分割共享编码器的方式利用语义信息实现深度估计增强的多任务训练，但仅仅通过共享编码器并不能充分利用语义信息. 因此，在如图4所示的对比方法深度估计结果可视化中，尤其是标注区域，可以看出相较于SGDepth, USegDepth捕获了更多的细粒度信息，特别是在物体边界，恢复出了更清晰的轮廓. 相较于没有联合语义信息，仅包含深度估计网络和位姿估计网络的Monodepth2[3]，整体轮廓模糊且细节信息丢失，因此，证明联合语义分割任务可以增强深度估计表现.

3.5 消融实验

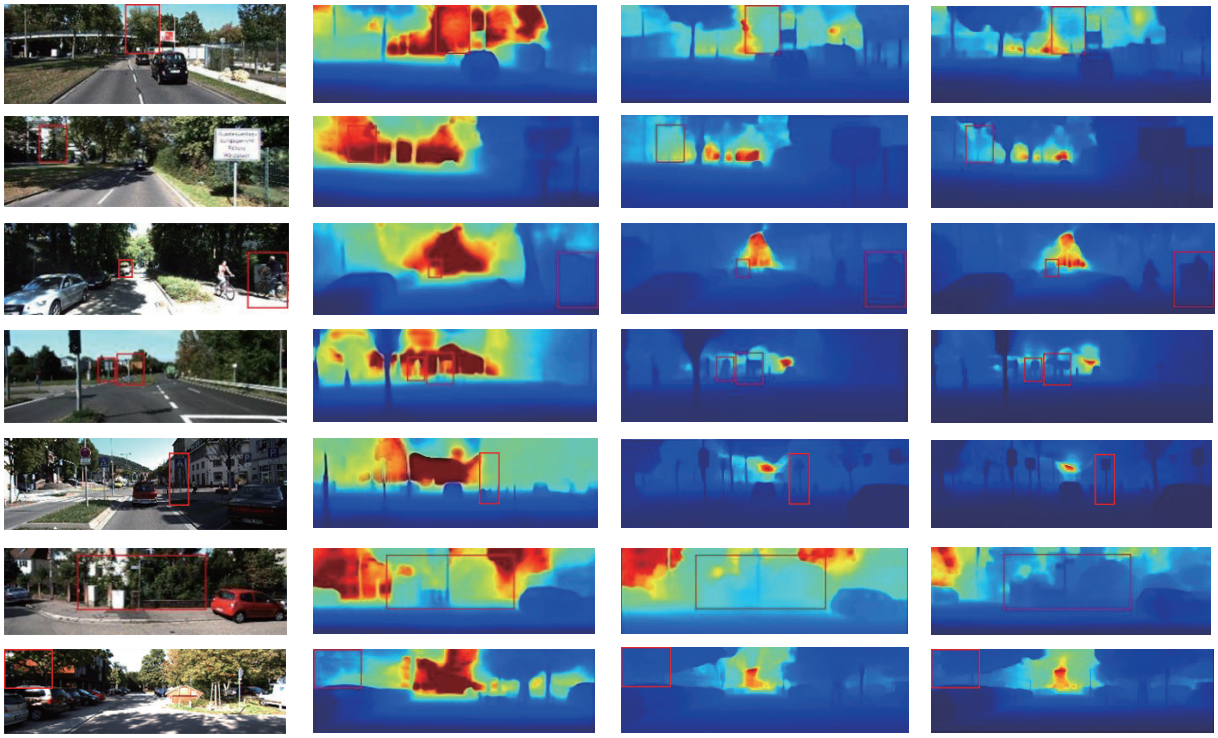
表2评估了每一个策略的有效性，其中Baseline代表设置共享编码器为ResNet50网络，损失函数仅包含光度损失和平滑损失的框架. 首先通过联合语义分割任务，采用共享编码器的方式共同训练，添加交叉熵损失L_{ce}作为监督信号，有效提升模型精度，证明语义信息有益于深度估计性能的提升. 然后，单独引入MTFEModule构建共享编码器，通过细化特征

Table 1 Comparison Results of Depth Estimation on KITTI Datasets

表 1 数据集 KITTI 上的深度估计对比结果

模型方法	语义模块	错误率				准确率		
		AbsRel	*SqRel*	*RMSE*	*RMSElog*	δ_1	δ_2	δ_3
Zhou 等人[7]		0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhao 等人[25]		0.146	1.084	5.445	0.221	0.807	0.936	0.976
Gordon 等人[26]	√	0.128	0.959	5.230	0.212	0.845	0.947	0.976
SceneNet[27]	√	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Lee 等人[28]	√	0.112	0.777	4.772	0.191	0.872	0.959	0.982
Zou 等人[29]		0.115	0.871	4.778	0.191	0.874	0.963	0.984
Guizilini 等人[30]	√	0.113	0.831	4.663	0.189	0.878	0.971	0.983
SGDepth[31]	√	0.112	0.833	4.688	0.190	0.884	0.961	0.981
Monodepth2[3]		0.110	0.831	4.642	0.187	0.883	0.862	0.982
PackNet-SfM[32]		0.111	0.785	4.601	0.189	0.878	0.960	0.982
SAFENet[33]	√	0.112	0.788	4.582	0.187	0.878	0.963	0.983
HR-Depth[34]		0.109	0.792	4.632	0.185	0.884	0.962	0.983
VTDepthB2[35]		0.105	0.762	4.530	0.182	0.893	0.964	0.983
MonoFormer[36]		0.104	0.846	4.580	0.183	0.891	0.962	0.982
USegDepth（本文）	√	**0.102**	**0.657**	**4.328**	**0.175**	**0.896**	**0.968**	**0.984**

注：“√”表示该方法联合了语义信息，黑体值表示最优结果.



(a) 输入图片 (b) Monodepth2[3] (c) SGDepth[31] (d) USegDepth

Fig. 4 Visualization of depth estimation results on NYU-Depth dataset

图 4 NYU-Depth 数据集上的深度估计结果可视化

映射，进一步提升深度估计表现，证明了 MTFEModule 的有效性. 又通过单独引入语义引导的跨任务交互模块 CTIModule，赋予深度预测更多的语义感知，进一步改进基线. 最后，所有策略的共同作用显著提升了 USegDepth 的表现.

表 3 将是否融入 CTIModule 生成的深度估计结

Table 2 Results of Ablation Experiments on KITTI Datasets

表 2 数据集 KITTI 上的消融实验结果

模型方法	错误率				准确率		
	AbsRel	*SqRel*	*RMSE*	*RMSElog*	δ_1	δ_2	δ_3
Baseline	0.123	0.898	4.873	0.216	0.870	0.959	0.982
Baseline+L_{ce}	0.119	0.834	4.762	0.192	0.879	0.961	0.983
Baseline+L_{ce}+MTFEModule	0.108	0.783	4.613	0.189	0.882	0.962	0.983
Baseline+L_{ce}+CTIModule	0.106	0.754	4.571	0.186	0.885	0.963	0.983
USegDepth（本文）	**0.102**	**0.657**	**4.328**	**0.175**	**0.896**	**0.968**	**0.984**

注：黑体值表示最优结果.

Table 3 Experimental Results of Ablation with/Without CTIModule on Depth Estimation and Semantic Segmentation Tasks

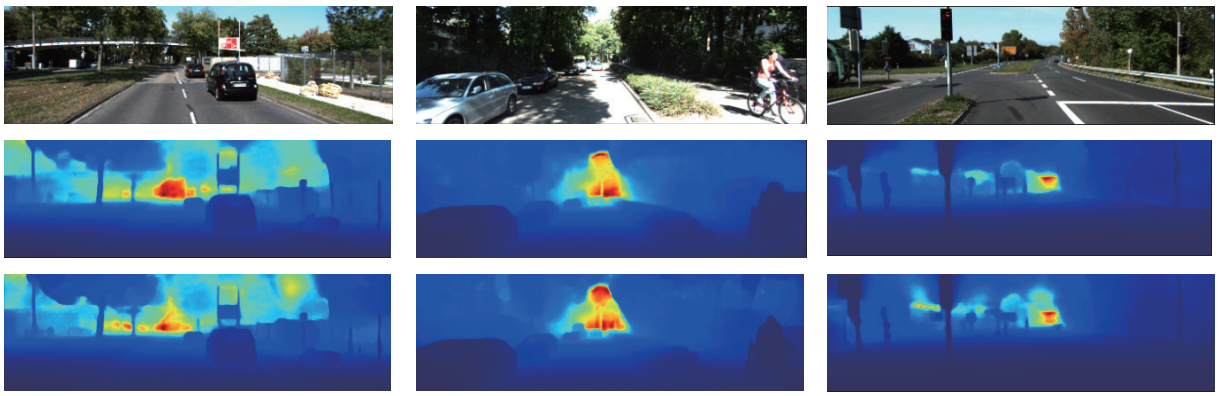
表 3 使用/不使用 CTIModule 在深度估计和语义分割任务上的消融实验结果

模型方法	语义分割	深度估计					
	MIoU	*AbsRel*	*SqRel*	*RMSElog*	δ_1	δ_2	δ_3
USegDepth w/o CTIModule	0.571	0.109	0.788	4.696	0.884	0.963	0.982
USegDepth（本文）	**0.577**	**0.102**	**0.657**	**4.328**	**0.896**	**0.968**	**0.984**

注：MIoU 为均交并比. 黑体值表示最优结果.

果及语义分割结果对比，证明 CTIModule 实现了语义分割和深度估计的双向增强. 为了更直观地对比分析，在图 5 中对深度估计结果进行可视化，其中第 1 行从 KITTI 数据集选取 RGB 样本图，第 2 行是模型在解码器部分不嵌入 CTIModule 时得到的深度估计图，第 3 行是模型在解码器部分嵌入 2 个 CTIModule 时得到的深度图. 可以看出，使用了 CTIModule 的模型恢复出了更清晰的边界以及更多的细节信息，说明该模块可以利用语义信息有效增强深度估计结果.

受多头自注意力启发，CTIModule 模块中采用多嵌入的形式，表 4 对嵌入数 H 的选择进行消融实验，结果表明采用多嵌入形式有效提升了深度估计表现.



(a) 样本图像1 (b) 样本图像2 (c) 样本图像3

Fig. 5 Visualization of ablation experiments with/without CTIModule module

图 5 使用/不使用 CTIModule 模块的消融实验可视化

Table 4 Results of Ablation Experiments on Number of Embeddings *H* for CTIModule Module

表 4 CTIModule 模块嵌入数 *H* 的消融实验结果

H	错误率				准确率		
	AbsRel	*SqRel*	*RMSE*	*RMSElog*	δ_1	δ_2	δ_3
1	0.107	0.784	4.616	0.184	0.890	0.965	0.983
2	0.105	0.759	4.571	0.181	0.889	0.966	0.984
4	**0.102**	**0.657**	**4.328**	**0.175**	**0.896**	**0.968**	**0.984**

注：黑体值表示最优结果.

4 结　论

本文提出了一种联合语义分割的精确单目深度估计方法. 为了充分利用语义信息, 深度估计任务与语义分割任务共享编码器, 该编码器由多任务特征提取模块堆叠构建, 其中采用了分组特征映射和多分支特征注意力融合策略, 旨在改善因感受野大小和缺乏跨通道的交互问题导致的特征表示能力下降问题, 实现了网络自适应地根据输入选择感受野大小, 提高了编码器的特征提取能力从而有效提升了模型精度. 在解码器各层之间嵌入跨任务交互模块, 利用来自深度估计特征图与语义分割特征图之间的交互信息, 生成丰富的空间细粒度表示, 细化特征, 实现双向特征增强. 通过在 KITTI 数据集的广泛评估, 证明本文所提的 USegDepth 模型方法具有较高的估计精度, 且相比于其他深度估计方法性能具有一定的提升.

作者贡献声明: 宋霄罡和胡浩越共同负责算法思路和实验方案设计; 胡浩越负责完成实验并撰写论文; 宁靖宇协助完成部分实验; 梁莉、鲁晓锋和黑新宏提出指导意见并修改论文.

参　考　文　献

[1] Jiang Junjun, Li Zhenyu, Liu Xianming. Overview of monocular depth estimation methods based on deep learning[J]. Chinese Journal of Computers, 2022, 45(6): 1276−1307(in Chinese)
(江俊君, 李震宇, 刘贤明. 基于深度学习的单目深度估计方法综述[J]. 计算机学报, 2022, 45(6): 1276−1307)

[2] Cheng Xinjing, Wang Peng, Yang Ruigang. Learning depth with convolutional spatial propagation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2361−2379

[3] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 6602–6611

[4] Imran S, Long Yunfei, Liu Xiaoming, et al. Depth coefficients for depth completion[C]//Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2019: 12438–12447

[5] Guo Xiaoyang, Li Hongsheng, Yi Shuai, et al. Learning monocular depth by distilling cross-domain stereo networks[C]//Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 484–500

[6] Luo Yue, Ren J, Lin M, et al. Single view stereo matching[C]//Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 155–163

[7] Shu Chang, Yu Kun, Duan Zhixiang, et al. Feature-metric loss for self-supervised learning of depth and egomotion[C]//Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2020: 572−588

[8] Yin Zhichao, Shi Jianping. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose[C]// Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 1983–1992

[9] Qi Xiaojun, Liao Renjie, Liu Zhengzhe, et al. GeoNet: Geometric neural network for joint depth and surface normal estimation[C]//Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 283–291

[10] Saha S, Obukhov A, Paudel P D, et al. Learning to relate depth and semantics for unsupervised domain adaptation [C]//Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2021: 8193−8203

[11] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Proc of the 28th Conf and Workshop on Neural Information Processing System (NIPS). Cambridge, MA: MIT, 2014: 2366–2374

[12] Zhou Tinghui, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 6612–6619

[13] Wang Peng, Shen Xiaohui, Lin Zhe, et al. Towards united depth and semantic prediction from a single image[C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2015: 2800–2809

[14] Jafari O H, Groth O, Kirillov A, et al. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation[C]//Proc of the 34th IEEE Int Conf on Robotics and Automation (ICRA). Piscataway, NJ: IEEE, 2017: 4620–4627

[15] Liu Beyang, Gould S, Koller D. Single image depth estimation from predicted semantic labels[C]//Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2010: 1253–1260

[16] Zhu Shenjie, Brazil G, Liu Xiaoming. The edge of depth: Explicit constraints between segmentation and depth[C]//Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 13116–13125

[17] Zhang Yuliang, Zhao Zhilong, Fu Weiping, et al. Monocular depth estimation based on edge Semantic information[J]. Science Technology and Engineering, 2022, 22(7): 2761−2769(in Chinese)
(张玉亮, 赵智龙, 付炜平, 等. 融合边缘语义信息的单目深度估计[J]. 科学技术与工程, 2022, 22(7): 2761−2769)

[18] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Proc of the 29th Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2015: 2017–2025

[19] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment:

From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600−612

[20] Park K, Patten T, Vincze M. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation[C]//Proc of the 17th IEEE Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2019: 7668–7677

[21] Jung H, Park E, Yoo S. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation [C]//Proc of the 18th IEEE Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2021: 12642–12652

[22] Godard C, Aodha O M, Firman M , et al. Digging into self-supervised monocular depth estimation[C]//Proc of the 17th IEEE Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2019: 3827−3837

[23] Zhu Yi, Sapra K, Reda F A, et al. Improving semantic segmentation via video propagation and label relaxation[C]//Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2019: 8856–8865

[24] Menze M, Geiger A. Object scene flow for autonomous vehicles[C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2015: 3061–3070

[25] Zhao Chaoqiang, Gary G Y, Sun Qiyu, et al. Masked GAN for unsupervised depth and pose prediction with scale consistency[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5392−5403

[26] Gordon A, Li H, Jonschkowski R, et al. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2019: 8976–8985

[27] Chen P Y, Liu A H, Liu Yencheng. et al. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation[C]//Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2019: 2619–2627

[28] Lee S, Im S, Lin S, et al. Learning monocular depth in dynamic scenes via instance-aware projection consistency[C] //Proc of the 35th AAAI Conf on Artificial Intelligence (AAAI). Palo Alto, CA: AAAI, 2021: 1863–1872

[29] Zou Yuliang, Ji Pan, Tran Q H, et al. Learning monocular visual odometry via self-supervised long-term modeling[C]//Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2020: 710–727

[30] Guizilini V, Hou Rui, Li Jie, et al. Semantically-guided representation learning for self-supervised monocular depth[C/OL]//Proc of the 8th Int Conf on Learning Representations (ICLR). Washington: ICLR, 2020[2023-12-26]. https://iclr.cc/virtual_2020/poster_ByxT7TNFvH.html

[31] Klingner M, Termöhlen J A, Mikolajczyk J, et al. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance[C]//Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2020: 582–600

[32] Guizilini V, Ambrus R, Pillai S, et al. 3D packing for self-supervised monocular depth estimation[C]//Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 2485−2494

[33] Choi J, Jung D, Lee D, et al. SaFENet: Self-supervised monocular depth estimation with semantic-aware feature extraction[J]. arXiv preprint, arXiv: 2010.02893, 2020

[34] Lyu Xiaoyang, Liu Liang, Wang Mengmeng, et al. HR-Depth: High resolution self-supervised monocular depth estimation[C]//Proc of the 35th AAAI Conf on Artificial Intelligence (AAAI). Palo Alto, CA: AAAI, 2021: 2294–2301

[35] Karpov A, Makarov I. Exploring efficiency of vision transformers for self-supervised monocular depth estimation[C]//Proc of the 25th IEEE Int Symp on Mixed and Augmented Reality (ISMAR). Piscataway, NJ: IEEE, 2022: 711–719

[36] Bae J, Moon S, Im S. Deep digging into the generalization of self-supervised monocular depth estimation[C]//Proc of the 36th AAAI Conf on Artificial Intelligence (AAAI). Palo Alto, CA: AAAI, 2022: 187–196



Song Xiaogang, born in 1987. PhD, associate professor. His main research interests include autonomous navigation of unmanned aerial vehicles, computer vision, and image processing.

宋霄罡, 1987 年生. 博士, 副教授. 主要研究方向为无人机自主导航、计算机视觉、图像处理.



Hu Haoyue, born in 1997. Master. Her main research interests include computer vision, image processing, and depth estimation.

胡浩越, 1997 年生. 硕士. 主要研究方向为计算机视觉、图像处理、深度估计.



Ning Jingyu, born in 1997. Master candidate. His main research interests include depth estimation and semantic segmentation.

宁靖宇, 1997 年生. 硕士研究生. 主要研究方向为深度估计、语义分割.



Liang Li, born in 1964. Master, associate professor. Her main research interests include computer vision and image processing, embedded computer control, and intelligent control.

梁　莉, 1964 年生. 硕士, 副教授. 主要研究方向为计算机视觉与图像处理、嵌入式计算机控制、智能控制.



Lu Xiaofeng, born in 1976. PhD, professor. His main research interests include computer vision, intelligent computing, and pattern recognition.

鲁晓锋, 1976 年生. 博士, 教授. 主要研究方向为计算机视觉、智能计算、模式识别.



Hei Xinhong, born in 1976. PhD, professor. His main research interests include computer vision and artificial intelligence.

黑新宏, 1976 年生. 博士, 教授. 主要研究方向为计算机视觉、人工智能.