文章编号:1671-4598(2026)06-0199-09 DOI:10.16526/j.cnki.11-4762/tp.2026.06.025 中图分类号:TP391.41 文献标识码:A

基于切片自查询的室外自监督单目深度估计

张 刚[1],张晖敏[2],张培媛[2],张海龙[3]
(1.江苏建筑职业技术学院 新能源工程学院,江苏 徐州 221116;
2.中国矿业大学 信息与控制工程学院,江苏 徐州 221116;
3.中国矿业大学 计算机科学与技术学院,江苏 徐州 221116)

摘要:针对室外单目深度估计中垂直方向敏感与局部卷积感受野受限导致的全局深度信息缺失问题,开展列方向全局建模研究;在编码阶段引入垂直像素自注意力以提取列级全局深度信息;在解码路径中对U-Net视觉特征执行按列重标定,使含全局先验的列获得更高权重并抑制噪声列;同时设置自查询模块,对编码器与解码器输出特征进行粗粒度查询,构建自生成代价体与深度分布特征并加权聚合以回归深度;在KITTI基准上,*AbsRel*降至0.087,*SqRel*降至0.653,*RMSE*降至4.120,$\delta<1.25$提升至0.921;在Make3D上,*AbsRel*由0.306降至0.304,*RMSE*由6.856降至6.778,*RMSE*log由0.151降至0.149;结论显示,列方向全局依赖的显式建模与概率融合弥补了局部卷积的全局感知不足,提升远距与纹理稀疏区域的估计稳定性与跨域鲁棒性,能够满足复杂室外场景对单目深度估计精度与泛化能力的需求。

关键词:深度学习;深度估计;自监督学习;注意力机制

Outdoor Self-Supervised Monocular Depth Estimation Based on Slice Self Query

ZHANG Gang[1], ZHANG Huimin[2], ZHANG Peiyuan[2], ZHANG Hailong[3]
(1. School of New Energy Engineering, Jiangsu Vocational Institute of Architectural Technology, Xuzhou 221116, China;
2. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China;
3. College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: To reduce the loss of global depth information caused by vertical-direction sensitivity and limited receptive fields of local convolutions in outdoor monocular depth estimation, research is conducted on the column-wise global modeling approach. During the coding process, a vertical-pixel self-attention module is employed to extract column-level global information; During the decoding process, the visual features of U-Net are re-calibrated by column-wise, making global prior columns to obtain higher weights and suppress noisy columns. Also, a self-query module is set up to perform coarse queries on the output features of the encoder and decoder, constructing a self- generated cost volume and depth-distribution features, thus performing weighted aggregation for depth regression. On the KITTI, the *AbsRel* decreases to 0.087, the *SqRel* to 0.653, the *RMSE* to 4.120, and δ increases from less than 1.25 to 0.921; On the Make3D, The *AbsRel* decreases from 0.306 to 0.304, the *RMSE* from 6.856 to 6.778, and the *RMSE*log from 0.151 to 0.149. The results show that the column-wise global dependency modeling and probabilistic fusion compensate for the lack of global perception in local convolutions, enhancing estimation stability and cross-domain robustness in long-range and sparse texture regions, thereby meeting the requirements of complex outdoor scenarios for monocular depth estimation accuracy and generalization capability.

Keywords: deep learning; depth estimation; self-supervised learning; attention mechanisms

收稿日期:2025-07-07; **修回日期**:2025-08-17。
基金项目:国家重点研发计划项目(2021YFC2902701);徐州市基础研究计划项目(KC22058)。
作者简介:张 刚(1986-),男,博士,副教授。
引用格式:张 刚,张晖敏,张培媛,等.基于切片自查询的室外自监督单目深度估计[J].计算机测量与控制,2026,34(6):199-207.

0 引言

深度估计作为计算机视觉的核心任务之一，旨在从二维图像中推断出三维场景中的深度信息。在自动驾驶[1]、机器人导航[2]、虚拟现实[3]、增强现实[4]、三维重建[5]以及摄影和电影制作[6]等领域具有重要应用价值。在这些场景中，深度估计技术有助于机器更准确地理解场景结构，从而实现更高效、更安全的交互和操作。

在深度估计的训练方法中，自监督深度估计利用图像之间的几何关系和投影约束来构建自监督信号，从而避免对真实深度标签的依赖[7]。这使得自监督深度估计在训练过程中无需收集昂贵且精确的真实深度数据[8]，降低了数据获取和标注的成本。此外，自监督深度估计通过图像重建等任务来间接监督深度估计的学习过程，使得模型能够学习到更加丰富的深度信息和上下文关系，这有助于提高深度估计的精度和泛化性能。单目自监督深度估计技术在深度估计领域具有广泛的应用前景和独特的优势，但仍然存在一些挑战。首先，深度估计的精度和稳定性需要在各种场景条件下得到进一步提高。其次，计算效率和实时性能也是需要考虑的重要因素，特别是在对实时性要求较高的应用中。此外，如何处理遮挡、动态场景和光照变化等问题也需要进一步探索。

综上，面向室外自监督单目深度估计中的全局深度感知不足问题，提出一种基于切片自查询与切片概率融合的列向全局建模框架。前者在列维度显式建模长程相关获取列级全局先验，后者突出重要全局深度，抑制不重要的全局深度。在保持训练设置不变的条件下，于 KITTI 与 Make3D 上获得稳定改进，显示出全局信息的加入使得网络性能更优。

1 深度估计研究进展

深度估计算法可分为传统方法和深度学习方法两大类。深度学习方法进一步分为监督深度估计和自监督深度估计，其区别在于训练过程中是否需要真实深度信息进行监督。在深度学习的框架下，改进网络结构是提高深度估计性能的一种关键手段，注意力机制在改进网络结构中发挥着重要的作用。

1.1 传统方法

传统方法中，文献［9］提出了一种基于注意力融合代价体积的立体匹配方法，通过结合不同特征尺度的信息来优化视差估计，从而提高了深度估计的准确性。文献［10］提出了一种细节－语义协作网络，通过选择性注意力机制提取并融合图像的细节与语义信息，从而提高了对复杂场景深度的估计精度。文献［11］提出了一种简化的单目深度估计网络，通过引入多尺度特征融合和注意力机制，在保证计算效率的同时显著提高了深度预测精度。

1.2 监督深度估计

传统深度估计算法往往存在泛化能力弱、推理时间长、估计精确度低等问题。深度学习模型克服了传统方法的局限。单目深度估计是一个不适定问题，为了解决此问题，引入了局部预测、非参数场景采样、端到端的监督学习。文献［12］提出了一种多尺度深度估计网络，通过层次化 Transformer 结构捕获图像的上下文信息，并采用细节－语义协作机制增强特征表示，从而提高了深度估计的准确性。文献［13］提出了一种轻量级的深度估计网络，通过深度可分离卷积和 Transformer 结合的方法来捕获局部和全局特征，并引入多尺度卷积注意力模块以提高细节恢复能力。文献［14］提出了 Aligned-MTL 多任务学习框架，通过对联合深度估计与表面法线任务的梯度进行对齐，显著提升了多任务联合训练的稳定性和性能。文献［15］通常使用三维代价体积来表示不同深度假设下的匹配概率，从而实现简单而有效的深度估计。文献［16］提出了一种粗到细的双阶段训练策略，先在粗略阶段学习运动目标的深度，再在精细阶段进一步细化深度估计。

1.3 自监督深度估计

自监督深度估计分为双目深度估计和单目深度估计。文献［17］提出了 NeRF 监督的深度立体匹配方法，通过将 Stereo 网络的输出与 NeRF 渲染结果进行一致性重投影来优化视差估计，从而显著提升了深度预测精度。文献［18］提出了一种基于不确定性伪标签的自监督立体匹配方法，通过对像素级不确定性估计生成高置信伪标签，并结合几何一致性损失来增强模型在遮挡和低纹理区域的匹配鲁棒性。文献［19］提出了一种联合视差与不确定性估计的新型损失，通过软直方图对视差误差分布进行建模，并将其与不确定性分布匹配，以在分类式深度平面网络中提升预测精度和可靠性。文献［20］提出了 Context-Enhanced Stereo Transformer，通过在 Transformer 编码器中融入多尺度上下文与代价体积特征，有效增强了视差图的细节与连续性。文献［21］采用平面视差几何，将连续帧中的像素对应关系分解为平面单应性和视差残差，从而使重建过程更加稳定和准确。

相比于自监督双目深度估计，自监督单目深度估计不仅需要估计深度信息，还需要估计位姿信息，才能重建图像，重建图像和目标图像构成约束进而监督网络。单目深度估计常用的训练方式为视频帧训练，输入视频序列，输出深度图。文献［22］提出了 GasMono 框架，该方法首先通过多视角几何计算粗略相机位姿，然后在训练中优化这些位姿，并结合 Transformer 和自蒸馏机

制提高了深度估计的鲁棒性。文献［23］提出了一种多帧深度估计方法，该方法通过分离场景中的刚体运动和遮挡区域来补偿光度一致性损失，从而提升了在动态场景中的深度估计精度。文献［24］提出了一种层次化细分类网络，通过将深度估计任务转化为多级分类问题，提升了模型在不同深度范围内的预测精度和泛化能力。文献［25］提出利用生成式扩散模型中学到的丰富先验进行深度估计，以显著提升深度模型的泛化能力。此外，扩展卡尔曼滤波模型[26]、3D 点云重建[27]、马尔可夫随机场[28]、利用光流估计处理遮挡等方法的推广使单目深度估计方法与双目估计方法性能差异不断缩小，但是在边缘细节、噪声处理、估计精度等方面仍有一定提升空间。

2　方法

室外数据集在垂直列像素上，深度存在差异，水平像素上深度值差异较小。为了充分使用方向敏感性这一性质，本文设计了两个模块，切片自查询层和切片概率融合层。单目深度估计常规做法是使用 Unet 提取图像特征，这些特征并没有关注垂直像素上深度的变化。切片自查询层将特征按列分割，计算每一列之间的相关性，列特征是深度信息的全局特征，对列向量进行切片自查询，进而融合全局信息。之后使用小型 Vision Transformer 提取粗粒度深度信息，使用粗粒度深度信息对 Unet 输出的深度图融合查询，从而提升网络性能。

2.1　深度估计与图像重建

深度估计的任务是输入一张图像，输出深度图。对于深度学习模型而言，深度估计的任务为学习一个函数，该函数描述了输入图像和输出深度图的关系，图像输入深度学习模型可以直接计算出其对应的深度图。自监督深度估计模型中包含深度估计以及位姿估计两部分，深度估计模型估计出当前帧的深度图，位姿网络根据当前相邻的两帧估计出相机位姿，联合深度图和相机位姿可以重建出另一个视角下的图像。如图 1 所示，以输入左视图为例，左视图经过深度估计网络得出预测深度值，如公式（1）所示：

$$\hat{d}(u,v) = \mathrm{model}[I^l(u,v)] \tag{1}$$

其中：(u, v) 为左视图 I^l 的像素坐标，$\hat{d}$ (u, v) 为 (u, v) 像素坐标对应的预测深度，$(u, v, \hat{d})$ 组成三维坐标 (X, Y, Z)。三维坐标 (X, Y, Z) 经过透视投影模型可以将其投影到二维平面的像素坐标 (u, v)。透视投影模型计算方式如公式（2）所示：

$$(u,v,1) = \boldsymbol{K}(\boldsymbol{R},\boldsymbol{t})(X,Y,Z,1) \tag{2}$$

其中：$\boldsymbol{K}$ 为相机内参矩阵，$(\boldsymbol{R}, \boldsymbol{t})$ 为外参矩阵，$\boldsymbol{R}$ 为旋转矩阵，$\boldsymbol{t}$ 表示平移矩阵，(X, Y, Z) 是三维坐标，二维平面像素坐标和三维坐标中的 1 是为了便利矩阵形式的表示而添加进去的。左视图的像素坐标 I^l (u, v) 输入深度估计网络得到预测深度值 $\hat{d}$ (u, v)，像素坐标 I^l (u, v) 和预测深度值 $\hat{d}$ (u, v) 组成左视图视角下的三维坐标 (X, Y, Z)，左视图 I^l 和下一帧图像 I_t^l 输入位姿估计网络，预测出旋转矩阵 $\boldsymbol{R}$ 和平移矩阵 $\boldsymbol{t}$，在透视投影模型的作用下，最终得到下一帧图像视角下的预测右视图 $\hat{I}^r$。在已知深度信息和位姿的条件下，左视图可以重建出右视图，重建右视图和原始右视图比较相似，至此形成自监督深度估计。

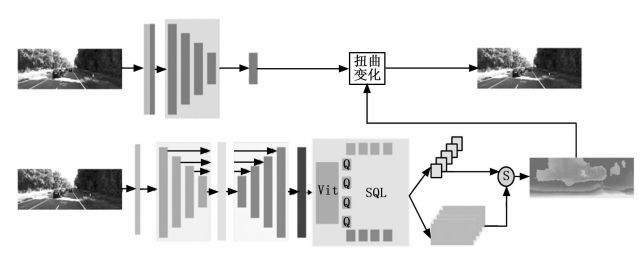

图 1　自监督单目深度估计架构

2.2　提取视觉特征

输入一张 RGB 彩色图像，尺寸为 $c \times h \times w$，经过卷积神经网络之后得到特征 l，之后特征 l 进入切片自查询层，聚焦垂直方向上的全局信息。卷积神经网络，由于使用小卷积，感受野小，无法提取全局信息，使用切片自查询层弥补了卷积网络无法提取全局信息的缺陷。在提取视觉特征阶段使用 Unet 网络，先经过 4 层下采样，之后进行 4 层上采样，在上采样的过程中使用跳跃连接，避免上采样过程中信息的丢失。在下采样提取特征的过程中，使用卷积神经网络和切片自查询层，卷积神经网络提取视觉特征，切片自查询层关注全局深度信息。

2.3　切片自查询层

在室外场景中，像素沿垂直方向通常对应由近到远的连续深度过渡。然而局部卷积的有限感受野难以捕获这一路径上的长程依赖，容易在远距与纹理稀疏区域产生不稳定估计。为弥补该不足，引入面向“列”的全局相关建模，使编码特征在列维度显式聚合全局深度线索。

如图 2 所示，经过卷积神经网络之后，输出的特征为 l，若 l 的尺寸 $c_1 \times h_1 \times w_1$，在输入切片自查询层之前，对 l 按列分割，每一列的尺寸为 $c_1 \times h_1$，一共列 w_1。每一列乘以变化矩阵 $\boldsymbol{W}_i^q$，$\boldsymbol{W}_i^k$，$\boldsymbol{W}_i^v$，得到对应的 q_i，k_i，v_i，q_i 查询所有的 k，即 q_i 乘以 $[k_1, k_2, k_3, \cdots, k_{w1}]$，得到 l_i 对应的 $[a_{i,1}, a_{i,2}, a_{i,3}, \cdots, a_{i,w1}]$ 查询分数，查询分数 $[a_{i,1}, a_{i,2}, a_{i,3}, \cdots, a_{i,w1}]$ 与 $[v_1, v_2, v_3, \cdots, v_{w1}]$ 进行点乘，点乘的结果即为 l_i 对应的垂直像素自注意力机制 b_i。查询分数代表着 l_i 与其他列之间的相关性，即不同列代表的全局深度信息之间的关系，不同的列之间相互查询，更好的提取全局深

度信息。$[l_1, l_2, l_3, \cdots, l_{w1}]$ 经过切片自查询之后分别得到 $[b_1, b_2, b_3, \cdots, b_{w1}]$，最后将所有 b_i 按列拼接成一个张量，得到列级全局先验 B。

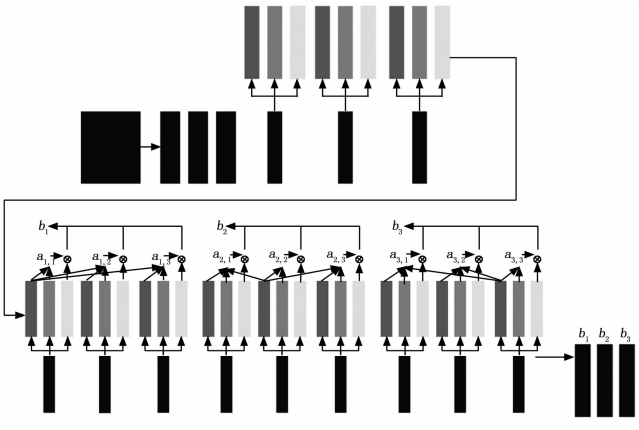

图 2 切片自查询层

上述列间自注意力在列维度聚合长距离结构的相关性，可为远距与纹理稀疏区域提供稳定的全局先验，弥补局部卷积在大范围依赖建模上的不足，并通过残差形式保持对基线特征的兼容。为了与后续的切片概率融合层衔接，网络以作为列级先验，与解码路径的列特征结合学习列权重，从而在远距与复杂结构处获得更稳定的一致性深度估计。由此得到的列级全局先验 B 作为切片概率融合层的输入，用于生成列权 p_j 并对解码器列特征重标定，实现先建模全局相关，再按重要性重加权的流程。

2.4 切片概率融合层

切片自查询层以编码器/解码器的特征按列进行聚合与相关性建模，输出列级全局深度描述子 b_j。该描述刻画了沿垂直方向的长程依赖与透视几何约束，弥补局部卷积感受野的不足，作为后续融合的全局深度先验输入到切片概率融合层。

切片概率融合层将来自当前尺度的列特征 s_j 与先验 b_j 进行拼接，经轻量两层感知器得到介于 0～1 的列权重 p_j，并在高度维度广播以对整列特征进行重标定，得到加权后的特征 S' 输入深度回归头。该先建模全局相关，再按重要性重加权的流程，使结构清晰、几何一致性更高的列获得更大权重，而纹理稀疏、反射/遮挡干扰明显的列被抑制，从而在近/中/远距均提升深度估计稳定性与精度，尤其对远距区域与细长结构更为有效；同时保持计算开销低，便于与多尺度解码器对接与复用。

尺寸为 $c\times h\times w$ 的彩色图像输入 Unet 网络，输出尺寸为 $c_1\times\frac{h}{2}\times\frac{w}{2}$ 的视觉特征 s，如图 3 所示。视觉特征 s 经过切片概率融合层，聚焦对深度信息影响大的列。Unet 提取的即时视觉特征没有区分不同列的重要程度，列像素包含全局深度信息，切片概率融合层计算每一列对应的权重，在全局深度信息上区分不同列。设输入特征 s 的尺寸为 $c_2\times h_2\times w_2$，首先将 s 按列分割，得到 w_2 列的 $c_2\times h_2$ 的特征，列特征包含着深度全局信息。对分割后的列特征进行全局平均池化，得到尺寸为 $w_2\times 1\times 1$ 的压缩特征 s_1，压缩特征 s_1 包含着全局深度信息，特征 s_1 经过两次全连接，第一次全连接将特征压缩至 $\frac{w_2}{16}\times 1\times 1$，第二次全连接将特征恢复至 $w_2\times 1\times 1$，得到垂直列像素的权重 p，最终视觉特征乘以权重得到最终的视觉特征。切片概率融合层计算每个垂直列特征的重要程度，将垂直像素权重乘以列特征，抑制不具有深度全局信息的列，加强包含具有深度全局信息的列。

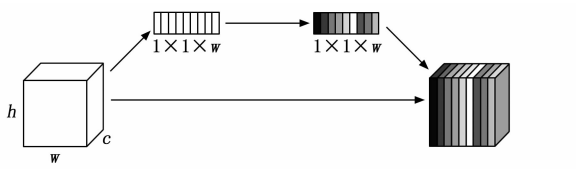

图 3 概率融合层

2.5 粗粒度自查询层

切片自查询层输出的全局深度特征不仅直接用于后续的粗粒度自查询，还作为先验信息传递给切片概率融合层。在该层中，先验全局特征用于指导列权重的计算，使得融合过程能够优先关注全局特征显著的列。这种由全局特征到列权重的引导机制，有助于在列加权过程中减少无关信息干扰，从而提升最终深度估计的准确性。

经过切片概率融合层处理后，得到特征 s_2，设 s_2 的尺寸为 $c_2\times h_2\times w_2$。在输入 4 层小型 Vision Transformer 之前，进行卷积核大小为 $p\times p$，步长为 p 的卷积操作，得到尺寸为 $c_2\times\frac{h_2}{p}\times\frac{w_2}{p}$ 的特征映射 F，特征映射 F 融合位置编码信息组成 4 层小型 Vision Transformer 的输入。小型 Vision Transformer 生成一组尺寸为 $c_2\times Q$ 的粗粒度查询 Q，其中 Q 是一个超参数。最后将这些粗粒度查询应用于即时视觉特征 s_2，以获得尺寸为 $h\times w\times Q$ 的自代价体积 V。

得到自代价体积之后，还需要计算潜在深度来估计深度栈 b，计数过程看作是一个信息聚合过程，并使用 softmax 以及加权和两种基本的信息聚合操作来实现计数操作。具体来说，对于自成本体积 V 中的每个平面，首先应用逐像素的 softmax 将体积平面转换为逐像素的概率映射。然后使用这个特征图在 s_2 中执行逐像素视觉表示的加权和。经过这个过程，得到纬度为 C 的 Q 个向量，每个平面表示深度计数。最后，将它们连接起来，

并将其输入 MLP 以回归深度栈 b，如公式（3）所示：

$$b = MLP\left[\bigoplus_{i=1}^{Q}\sum_{(j,k)=(1,1)}^{(h,w)}\mathrm{softmax}(V_i)_{j,k}\cdot S_{j,k}\right] \tag{3}$$

上一步骤中使用一组粗粒度查询来执行逐像素查询，每个查询都代表不同的图像上下文并产生自己的深度估计。这一步骤结合所有这些深度估计来得到最终的深度图。首先，为了匹配形状 D 的深度箱 b 的尺寸，对自代价体积 V 应用 1×1 卷积得到深度平面集合。其次，应用一个面向平面的 softmax 操作将体积转换为面向平面的概率映射，如公式（4）所示：

$$p_{i,j,k} = \mathrm{softmax}(V)_{i,j,k},\quad 1\leqslant i\leqslant Q \tag{4}$$

最后，对于每个像素点，采用标准概率线性组合的方式对箱子的中心进行聚类计算深度，如公式（5）所示：

$$\tilde{d} = \sum_{i=1}^{N} c(b_i)\,p_{i,j,k},\quad 1\leqslant j\leqslant h, 1\leqslant k\leqslant w \tag{5}$$

其中：$c(b_i)$ 为第 i 个 bin 的中心深度，如公式（6）所示：

$$c(b_i) = d_{\min} + (d_{\max} - d_{\min})\left(\frac{b_i}{2} + \sum_{j=1}^{i-1} b_j\right) \tag{6}$$

2.6　损失函数

损失函数由两部分组成，光度损失函数和边缘平滑损失函数。在现实场景中，静止摄像机和移动物体等情况会打破移动摄像机和静态场景的假设，损害自监督深度估计器的性能。为了提高深度预测的准确性，在光度损失和边缘平滑损失上加入运动物体的掩膜，借此去除运动物体对深度估计影响。

光度损失函数 PE 使用了 L_1 范数和 SSIM[29]，光度损失函数的计算方式如公式（7）所示：

$$PE(I_a, I_b) = \frac{\alpha}{2}[1 - SSIM(I_a, I_b)] + (1-\alpha)\left\|I_a - I_b\right\|_1 \tag{7}$$

边缘感知平滑损失的功能为正则化无纹理区域的深度，计算方式如公式（8）所示：

$$L_s = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|} \tag{8}$$

运动物体掩码[30]的计算方式如公式（9）所示：

$$\mu = \left[\min_{t'} PE(I_t, I_{t'\to t}) \leqslant \min_{t'} PE(I_t, I_{t'})\right] \tag{9}$$

最终的损失函数结合了光度损失函数和边缘平滑损失函数，使用运动掩码去除运动物体，计算方式如公式（10）所示：

$$L = \mu L_p + \lambda L_s \tag{10}$$

3　实验结果与分析

KITTI 数据集是一个广泛用于自动驾驶和计算机视觉研究的公开数据集，该数据集主要包含在城市环境中以不同天气和不同时间采集的传感器数据，例如立体视觉、激光雷达和惯性导航系统数据。数据集里面有驾驶视频以及对应的深度信息，训练深度估计网络使用该部分的数据集。数据集的分离使用 Eigeen 的方法，在使用数据集训练测试之前，需要剔除视频中的静态帧。剔除静态帧使用 zhou 的方法。数据集中 39 810 张图片用于训练，4 424 张图片用于评估。所有的图像使用了相同的相机内参，相机主点设置为图像中点，焦距设置为 KITTI 所有焦距的平均值。在模型评估时，设置了深度最大值 80，由于预测深度和真实深度之间存在缩放关系，预测深度采用相对深度表示。网络训练时使用 RTX3090 显卡，深度学习框架使用 Pytorch 框架，训练时输入图像的尺寸为 192×640，优化器使用 Adam 优化器，一阶动量参数 β_1 设置为 0.9，二阶动量参数 β_2 设置为 0.999，学习率设置为 10^{-4}。损失函数中 SSIM 的超参数设置为 0.85，平滑损失中的设置为 10^{-3}。训练完成后，采用绝对相对误差（AbsRel）、均方根误差（RMSE）、均方根相对误差（SqRel）3 个误差指标与 3 个不同阈值的精确度指标对深度估计结果进行测评，如表 1 所示。

相对误差：

$$AbsRel = \frac{1}{N}\sum\frac{\left|d_i - d_i^*\right|}{d_i} \tag{11}$$

均方根误差：

$$RMSE = \sqrt{\frac{1}{N}\sum\left|d_i - d_i^*\right|^2} \tag{12}$$

表 1　不同深度估计算法性能比较

Method	Train	*AbsRel* ↓	*SqRel* ↓	*RMSE* ↓	*RMSE*log ↓	$\delta<1.25$ ↑	$\delta<1.25^2$ ↑	$\delta<1.25^3$ ↑
文献[31]	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
文献[32]	M	0.106	0.861	4.699	0.185	0.889	0.962	0.982
文献[30]	MS	0.106	0.818	4.750	0.196	0.874	0.957	0.979
文献[33]	M	0.106	0.799	4.662	0.187	0.889	0.961	0.982
文献[34]	M	0.105	0.769	4.535	0.181	0.892	0.964	0.983
文献[35]	M	0.096	0.720	4.458	0.175	0.897	0.964	0.984
文献[36]	M	0.090	0.713	4.261	0.170	0.914	0.966	0.983
文献[37]	MS	0.088	0.697	4.175	0.167	0.919	0.969	0.984
本文	MS	0.087	0.653	4.120	0.165	0.921	0.97	0.984

表 2 本文模块消融实验

切片自查询层	切片概率融合层	*AbsRel*	*SqRel*	*RMSE*	*RMSE*log	$\delta<1.25$ ↑	$\delta<1.25^2$ ↑	$\delta<1.25^3$ ↑
—	—	0.088	0.697	4.175	0.167	0.919	0.969	0.984
√	—	0.087	0.650	4.129	0.165	0.918	0.969	0.984
—	√	0.088	0.687	4.175	0.167	0.919	0.969	0.984
√	√	0.087	0.653	4.12	0.165	0.921	0.97	0.984

均方根相对误差：

$$SqRel = \frac{1}{N}\sum \frac{|\ d_i - d_i^*\ |^2}{d_i} \tag{13}$$

修正的精度指标：

$$\delta_k = \sum \max\left(\frac{a_i}{t_i}\ \frac{t_i}{a_i}\right) < \left(1+\frac{k}{100}\right) \tag{14}$$

3.1 对比实验

表 1 是各种深度估计算法性能的比较，误差指标上，本文算法误差比大多数算法误差小[30-37]；精度指标上，指标较之其它模型均有提升。从表 1 中的实验数据可以看出，相对误差下降 0.001，均方根相对误差下降 0.044，均方根误差下降 0.055，$\delta<1.25$ 指标提升 0.002。误差指标和精确度指标显示了本方法的有效性。

图 4 是不同深度估计算法在测试集上的深度图。Monodepth、Manydepth、SQLdepth 得出的深度图在边缘轮廓上比较模糊，本文算法在边缘上比较准确。在上图中的场景 1 中，Monodepth、Manydepth、SQLdepth 得出的深度图中两个行人的边界产生了重叠，本文算法中两个行人边界没有发生重叠；场景 2 中，Monodepth、Manydepth、SQLdepth 得出的深度中的汽车周围的深度存在散射现象，边缘轮廓平滑过渡，本文算法中汽车周围边缘清晰，不存在散射现象；在场景 4 中，Monodepth、Manydepth、SQLdepth 中汽车玻璃的深度存在偏差，本文算法中汽车边缘清晰；场景 5 中，Monodepth、Manydepth、SQLdepth 的路标深度一致性较差，路标杆深度不一致，本文算法不存在这个问题。

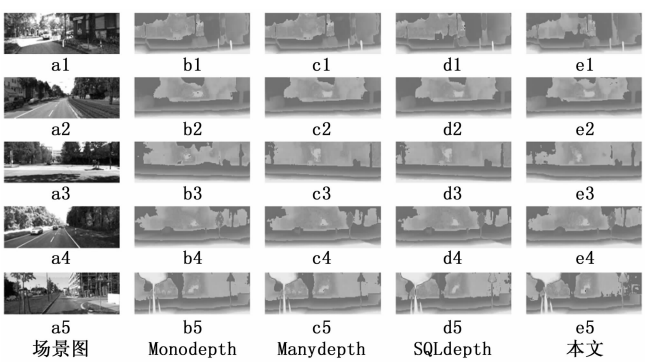

图 4 KITTI 数据集的可视化结果对比

3.2 消融实验

在神经网络编码层，使用切片自查询层机制，弥补卷积神经网络感受野小的问题，在保持小卷积的同时，关注深度全局信息。Unet 提取的即时视觉特征，在进行粗粒度查询之前，进行垂直像素注意力机制，计算不同列代表的深度信息的权重。消融实验旨在验证切片自查询层和切片概率融合层的有效性。本文的主干网络为 SQLdepth，在主干网络的基础上依次加入切片自查询层、切片概率融合层，以此验证模块的有效性。

从表 2 可以看出，本研究提出的模块对深度估计的指标都有不同程度的提升。切片自查询层融合深度全局信息，相对误差指标下降 0.001，均方根相对误差下降 0.047，均方根误差下降 0.046。切片概率融合层计算不同列的权重，均方根相对误差下降 0.010。同时使用两个模块，相对误差下降 0.001，均方根相对误差下降 0.044，均方根误差下降 0.055，精度指标 $\delta<1.25$ 提升 0.002。通过消融实验，可以证明本研究提出两个模块的有效性。切片自查询层通过在列方向建模长程相关，强化对环境中纵向连续结构的表达，减轻远距与弱纹理区域的过度平滑与轮廓粘连。切片概率融合层以列级权重对特征进行重标定，抑制由反射、遮挡和局部重建不一致引起的噪声，使近至中距离区域的 *SqRel* 与 *RMSE* 呈持续收敛趋势，整体上表现为 *AbsRel*、*SqRel*、*RMSE* 下降与 $\delta<1.25$ 提升的统一走向。

两模块协同时，切片自查询层提供的列级全局先验为概率融合层的权重学习提供依据，形成先建模全局相关，再按重要性重加权的流水线；结果是在近/中/远距离均获得更稳健的改进，其中远距段的结构保持与 $\delta<1.25$ 提升更为稳定，近—中距离的误差波动进一步减小，与表 2 的整体趋势一致。

图 5 是分别加入切片自查询层和切片概率融合层的实验效果图。从图中可以看出，加入模块之后深度图的边缘更加清晰。

3.3 泛化性实验

本研究中，神经网络的训练是在 KITTI 数据集上训练的，为了验证模型的泛化能力，使用 KITTI 数据集上的预训练模型，在不微调的情况下，使用 Make3D 数据集[38]上进行测试，测试的结果如表 3 所示，以文献 [37] 为主对比对象，并参其他方法用于横向对照，可以得出该方法的误差更低、精度更高，即视为跨域泛

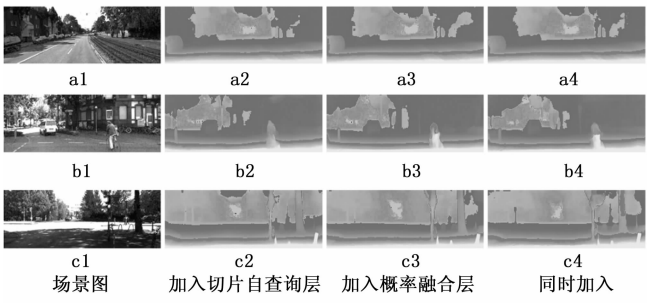

图 5　消融实验

化更强。

表 3　Make3D 数据集上的迁移测试结果

模型	训练方式	*AbsRel*	*SqRel*	*RMSE*	*RMSE*log
文献[5]	S	0.544	10.94	11.760	0.193
文献[22]	M	0.383	5.321	10.470	0.478
文献[39]	M	0.387	4.720	8.090	0.204
文献[30]	M	0.322	3.589	7.417	0.163
文献[34]	M	0.312	3.086	7.066	0.159
文献[37]	M	0.306	2.402	6.856	0.151
本文	MS	0.304	2.366	6.778	0.149

如图 6 所示，为该模型在 Make3D 数据集上推理结果，其中包含建筑立面、开阔道路、自然植被等多类场景。结合表 3 总体该方法最优的趋势，可做如下归因：在建筑立面与细长结构占优的片段，列方向全局相关的显式建模提升边缘保持与几何一致性，缓解轮廓粘连；在开阔道路与远端区域，对长程透视梯度的建模使远距误差收敛更稳定，$\delta<1.25$ 的提升更集中；在纹理稀疏或高反射表面，列级概率融合抑制由局部不一致带来的噪声，降低 *SqRel* 与 *RMSE* 波动。

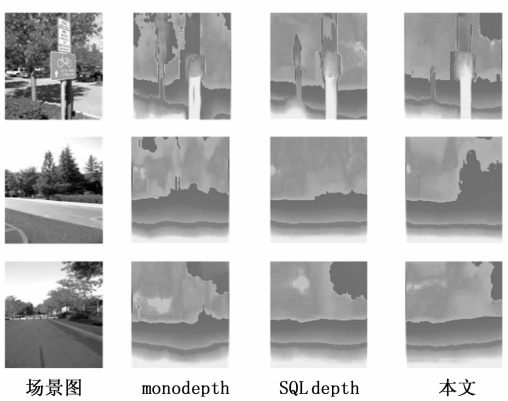

图 6　Make3D 数据集的可视化对比

4　结束语

本文提出一种垂直像素注意力的自监督单目深度估计，通过切片自查询，实现对全局深度信息的编码，小核卷积提取局部信息，切片自查询层提取全局信息，提取出特征之后，对每个列代表的全局深度进行概率融合，计算不同全局深度信息的权重。在 KITTI 上相较基线取得稳定改进；在 Make3D 的零微调迁移设置下，同样呈现小幅且一致的提升，表明列向全局建模与概率重标定的协同具有一定跨域有效性。后续工作将在夜间与恶劣天气下开展外部验证，并补充距离/条件分段评估；同时探索对照度变化与遮挡更鲁棒的重建约束与列向稀疏注意以提升跨域稳定性。

参考文献：

[1] 刘洋宏，付杨悠然，董性平. HDMapFusion：用于自动驾驶的多模态融合高清地图生成 [J/OL]. 计算机工程，2025：1－10.

[2] 杨　彪，王　狄，沈绍博，等. 基于深度测距的移动机器人自适应跟随 [J]. 计算机工程与设计，2024，45 (3)：896－903.

[3] 王　震，盖　孟，许恒硕. 基于虚拟现实技术的三维场景图像表面重建算法 [J]. 吉林大学学报 (工学版)，2022，52 (7)：1620－1625.

[4] 王振宇，许　驰，李　琳，等. 增强现实机器人的虚实同步手势交互方法 [J]. 计算机应用研究，2025，42 (8)：2290－2296.

[5] 李双全，张奇贤，姚海洋，等. 基于单目光学图像的水下场景三维重建技术 [J]. 激光杂志，2024，45 (11)：93－99.

[6] 程德强，张华强，寇旗旗，等. 基于层级特征融合的室内自监督单目深度估计 [J]. 光学精密工程，2023，31 (20)：2993－3009.

[7] 程德强，范舒铭，钱建生，等. 基于坐标感知注意的多帧自监督单目深度估计 [J]. 北京航空航天大学学报，2023：1－14.

[8] 宗长富，文　龙，何　磊. 基于欧几里得聚类算法的三维激光雷达障碍物检测技术 [J]. 吉林大学学报 (工学版)，2020，50 (1)：107－113.

[9] XU G，CHENG J，GUO P，et al. Attention concatenation volume for accurate and efficient stereo matching [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，New Orleans，2022：12981－12990.

[10] SONG Z，GAO Y，DAI S，et al. SIE-DepthNet：semantic-guided monocular depth estimation for dynamic environment [C] //International Conference on Extended Reality，Singapore，2024：195－210.

[11] LIU X，TANG S，FENG M，et al. A simple monocular depth estimation network for balancing complexity and accuracy [J]. Scientific Reports，2025，15 (1)：12860.

[12] SONG W，CUI X，XIE Y，et al. Monocular depth estimation via a detail semantic collaborative network for in-

door scenes [J]. Scientific Reports, 2025, 15 (1): 10990.

[13] SUI X, GAO S, XU A, et al. Lightweight monocular depth estimation using a fusion-improved transformer [J]. Scientific Reports, 2024, 14 (1): 22472.

[14] SENUSHKIN D, PATAKIN N, KUZNETSOV A, et al. Independent component alignment for multi-task learning [C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023: 20083 - 20093.

[15] LI R, GONG D, YIN W, et al. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes [C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023: 21539 - 21548.

[16] MOON J, BELLO J L G, KWON B, et al. From-ground-to-objects: coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior [C] //2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2024: 10519 - 10529.

[17] TOSI F, TONIONI A, De GREGORIO D, et al. Nerf-supervised deep stereo [C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023: 855 - 866.

[18] SHEN Z, SONG X, DAI Y, et al. Digging into uncertainty-based pseudo-label for robust stereo matching [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45 (12): 14301 - 14320.

[19] CHEN L, WANG W, MORDOHAI P. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation [C] //2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023: 17235 - 17244.

[20] LOU J, LIU W, CHEN Z, et al. Elfnet: Evidential local-global fusion for stereo matching [C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Paris, 2023: 17784 - 17793.

[21] YANG X, MA Z, JI Z, et al. Gedepth: ground embedding for monocular depth estimation [C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Paris, 2023: 12719 - 12727.

[22] ZHAO C, POGGI M, TOSI F, et al. Gasmono: geometry-aided self-supervised monocular depth estimation for indoor scenes [C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Paris, 2023: 16209 - 16220.

[23] FENG Z, YANG L, JING L, et al. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth [J]. European Conference on Computer Vision, 2022: 228 - 244.

[24] WANG Y, LIANG Y, XU H, et al. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation [C] //Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 5713 - 5721.

[25] KE B, OBUKHOV A, HUANG S, et al. Repurposing diffusion-based image generators for monocular depth estimation [C] //2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2024: 9492 - 9502.

[26] ZHANG S, ZHANG J, TAO D. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics [C] // European Conference on Computer Vision, Cham: Springer Nature Switzerland, 2022: 143 - 160.

[27] LIU Z, LI R, SHAO S, et al. Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33 (12): 7565 - 7577.

[28] 刘晓旻，杜梦珠，马治邦，等. 基于遮挡场景的光场图像深度估计方法 [J]. 光学学报，2020，40 (5): 0510002.

[29] WANG Z, BOVIK A S H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Trans on Image Process, 2004, 13 (4): 600 - 612.

[30] GODARD C, AODHA O M, FIRMAN M, et al. Digging into self-supervised monocular depth estimation [C] //2019 IEEE/CVF International Conference on Computer Vision, Seoul, 2019: 3827 - 3837.

[31] GUIZILINI V, AMBRUS R, PILLAI S, et al. 3D packing for self-supervised monocular depth estimation [C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020: 2482 - 2491.

[32] JOHNSTON A, CARINEIRO G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume [C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020: 4755 - 4764.

[33] ZHANG G, WANG J, LIU L, et al. Self-supervised joint learning framework of depth estimation via implicit cues [J]. ArXiv, 2020.

[34] YAN J, ZHAO H, BU P, et al. Channel-wise attention-based network for self-supervised monocular depth estimation [J]. International Conference on 3D Vision, 2021: 464 - 473.

[35] FENG Z，YANG L，JING L，et al. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth [J]. European Conference on Computer Vision，2022：228－244.

[36] WATSON J，AODHA O，PRISACARIU V，et al. The temporal opportunist：self-supervised multi-frame monocular depth [J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition，2021：1164－1174.

[37] WANG Y，LIANG Y，XU H，et al. SQLdepth：generalizable self-supervised fine-structured monocular depth estimation [J]. ArXiv，2023.

[38] SAXENA A，SUN M，NG A Y. Make3D：Learning 3D scene structure from a single still image [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2009，31 (5)：824－840.

[39] WANG C，BUENAPOSADA J，ZHU R，et al. Learning depth from monocular videos using direct methods [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，Salt Lake City，2017：2022－2030.

(上接第 88 页)

[12] 乌海能源. 乌海能源职工健康智能监测管理平台上线运行 [EB/OL]. (2024-03-27) [2025-11-01]. https://www.ceic.com/gjnyjtww/chnxgdt/202403/a39690023568 4dde99c9a9a06507871c.shtml.

[13] 李映辰，何宇斌，何锡祺，等. 基于云边协同的充电设施协调控制系统设计 [J]. 计算机测量与控制，2024，32 (6)：111－117.

[14] 陆俊杰，魏亚东，李晓峰，等. 基于 OCR 技术的航天器材料及器件试验数据识别系统 [J]. 计算机测量与控制，2023，31 (1)：282－288.

[15] KE G，MENG Q，FINLEY T，et al. LightGBM：a highly efficient gradient boosting decision tree [C] //Advances in Neural Information Processing Systems 30 (NIPS 2017)，Long Beach，CA：Curran Associates，Inc.，2017：3146－3154.

[16] EFFATI S，KAMARZARDI T A，AZIZI F F，et al. Web application using machine learning to predict cardiovascular disease and hypertension in mine workers [J]. Sci. Rep.，2024，14 (1)：31662.

[17] 童　兴. 高温煤矿作业人员热反应规律及热应激计算评价模型研究 [D]. 北京：中国矿业大学 (北京)，2018.

[18] 杨洪旭，崔风涛，刘　航，等. 皖北地区煤矿工人常见慢性病共病模式及相关性分析 [J]. 安徽预防医学杂志，2024，30 (6)：1－6.

[19] 付爱玲. 新疆煤矿工人职业紧张与高血压关系的双向队列及代谢组学研究 [D]. 乌鲁木齐：新疆医科大学，2024.

[20] 郭伟洁，植凯吉，白瑀皓，等. 基于身份证和人脸双重识别技术的智能门禁系统设计 [J]. 计算机测量与控制，2021，29 (2)：222－228.

[21] 彭庆国，唐再汐，王振宇，等. 一种基于体征监测的煤矿作业人员安全监测系统 [J]. 中国科技信息，2025 (8)：113－116.

[22] 耿仁亮. 辅助血压规范化测量的智能系统研究与应用 [D]. 南昌：南昌大学，2022.

[23] 杨思豪. 基于可穿戴手环多感知特征融合的身心健康评估方法研究 [D]. 成都：电子科技大学，2020.

[24] 张　璐，李冰鹃，段　特，等. 基于嵌入式与表情识别的人体生理参数监测系统设计 [J]. 计算机测量与控制，2025，33 (3)：45－53.

[25] 冯　晔，周瑾萱，谭　鑫，等. 基于 OpenCV 的运动控制与自动追踪系统设计与实现 [J]. 计算机科学与应用，2024，14 (2)：224－232.

[26] 李成勇，王　莎，陈成瑞. 基于 OpenCV 的人脸识别系统设计与实现 [J]. 国外电子测量技术，2021，40 (11)：168－172.

[27] 张杜娟，丁　莉，吴玉莲. 基于 Python＋Dlib 方法的人脸识别技术研究 [J]. 电子设计工程，2024，32 (1)：191－195.

[28] 邓垲镛，赵慧元，田　刚，等. 基于 LBPH 算法的人脸识别算法研究与设计 [J]. 中小企业管理与科技，2020 (3)：184－185.

[29] 卢　嫚，谢磊鑫. 基于 OpenCV 智能车牌及颜色识别 [J]. 自动化与仪表，2023，38 (8)：120－124.

[30] 张安然. 基于巡检机器人的数字仪表自动识别方法 [J]. 煤炭技术，2024，43 (3)：262－263.

[31] 邹　彤，乐　健，谌珺瑶，等. 自适应×字算法识别仪表数字 [J]. 仪表技术与传感器，2022 (4)：108－111.

[32] 潘宇强，姚　垚，张　林，等. 基于不规则目标检测的指针式仪表读数识别方法 [J]. 仪表技术与传感器，2024 (6)：100－105.

[33] 潘　桥，郑明杰. 云边端架构的工业产品质量检测系统设计 [J]. 计算机测量与控制，2025，33 (9)：63－73.

[34] 俞子舒，王一帆，曾　琛，等. 支持端边云多运行时协同应用的网程系统 [J]. 计算机研究与发展，2025，62 (12)：3042－3059.

[35] 郝振兴，康喜富，王基月，等. 基于物联网的液压泵站远程监控系统设计 [J]. 仪表技术与传感器，2022 (3)：80－83.

[36] 王志明，李　鹏，韦　杰，等. 基于边缘计算和多传感器融合的输电线路监测研究 [J]. 计算机测量与控制，2023，31 (11)：113－118.