李晗, 葛动元. 融合 HEALPix 网格和 Swin Transformer 的全景图像深度估计[J]. 智能计算机与应用,2026,16(8):18-25. DOI:10.20169/j.issn.2095-2163.25011701

融合 HEALPix 网格和 Swin Transformer 的全景图像深度估计

李 晗, 葛动元
(广西科技大学 机械与汽车工程学院, 广西 柳州 545006)

摘 要: 全景图像深度估计是一个具有广阔前景的研究方向。与二维平面图像相比,全景图像包含 360°×180°的视野,但其平面投影往往存在局部失真和不连续性,为解决这一问题,本文将 HEALPix(Hierarchical Equal Area isoLatitude Pixelization)网格与 Swin Transformer 相融合,提出一种新的全景图像深度估计模型。HEALPix 网格是一种分层等面积等纬度处理球面数据的方法,将整个球面的像素进行均匀地划分,还可以满足不同分辨率的需求。Swin Transformer 的自注意力机制和移动窗口能够捕捉像素之间的依赖关系和局部特征, HEALPix 网格的分层嵌套结构可以与窗口的移动和掩码的生成相融合。两者融合后可以对全景图像进行球面数据处理,有效地避免了全景图像平面投影的局部失真问题。最后,本文在 Stanford2D3D 数据集上进行测试,实验结果表明,在 MAE(0.142 9)、AbsRel(0.047 2)和 RMSE(0.097 27)等指标上表现均优于现有的基于深度学习的全景图像深度估计方法。

关键词: 全景图像; 深度估计; HEALPix 网格; 球面数据; Swin Transformer
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0018-08

Panoramic image depth estimation by integrating HEALPix grid and Swin Transformer

LI Han, GE Dongyuan
(School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, Liuzhou 545006, Guangxi, China)

Abstract: Panoramic image depth estimation is a research direction with broad prospects. Compared to 2D planar images, panoramic images cover a 360°×180° field of view, but their planar projections often suffer from local distortions and discontinuities. To address this issue, a new panoramic image depth estimation model is proposed by integrating the HEALPix(Hierarchical Equal Area isoLatitude Pixelization) grid with Swin Transformer. The HEALPix grid is a method for hierarchically equal-area iso-latitude pixelization of spherical data, which uniformly divides the entire sphere into pixels and can also meet different resolution requirements. The self-attention mechanism and moving windows of the Swin Transformer capture dependencies and local features between pixels. The hierarchical nested structure of the HEALPix grid can be seamlessly integrated with the movement of windows and the generation of masks. After integration, the combined model can process spherical data from panoramic images, effectively avoiding the local distortion issues caused by the planar projection of panoramic images. Finally, experiments were conducted on the Stanford2D3D dataset, and the results showed that the proposed method outperformed existing deep learning-based panoramic image depth estimation methods in terms of MAE (0.142 9), AbsRel (0.047 2), and RMSE (0.097 27).

Key words: panoramic image; depth estimation; HEALPix grid; spherical data; Swin Transformer

0 引 言

全景图像深度估计是深度估计领域的重要分支,是计算机视觉领域核心研究任务,其目标是获取全景图像对应的深度信息[1]。与传统二维图像相比,全景图像的视野更广,可完整地呈现 360°×180°的三维场景。全景图像不仅在虚拟现实、增强现实、自动驾驶、机器人导航等领域有着广泛应用。全景图像原生具备球面几何结构,与二维平面图像存在本质差异,在数据处理过程中需要将其投影到平面

基金项目: 国家自然科学基金(51765007)。
作者简介: 李 晗(2000—),男,硕士研究生。主要研究方向:全景视觉,深度估计。
通信作者: 葛动元(1970—),男,博士,研究员。主要研究方向:深度学习,相机标定。Email:gordon399@gxust.edu.cn。
收稿日期: 2025-01-17

上[2]，但常见的平面投影存在失真和不连续性，因此对全景图像进行深度估计时要面临几何失真和局部扭曲等问题。

传统获取图像深度信息的方法，通常依赖于特定的硬件设备，如通过激光雷达、结构光和双目相机等。激光雷达获得的深度信息精度高、范围广，但其成本较高[3]。结构光通过接收反射的光束来计算物体的深度信息，容易受到视距、光照和场景复杂性等因素影响。双目相机使用两个相机拍摄同一场景的不同视角图像，结合相机的内外参数来估算目标深度[4]。

随着深度学习在计算机视觉应用的发展，近年来涌现出大量基于深度学习的全景图像深度估计方法，这类方法可自动从大规模数据中学习到有效特征，显著提升了深度估计的精度和鲁棒性。Sun 等[5]提出了 HoHoNet 网络，该网络以特征金字塔为骨干网络，沿列向量进行编码，使用逆离散余弦变换将潜在向量展开到二维空间域，从而实现全景深度估计。考虑到全景图像 ERP 投影（Equirectangular Projection，ERP）的畸变问题，Zioulis 等[6]提出了 OmniDepth 网络，一种完全卷积的编码器-解码器网络，该网络使用可变卷积来处理畸变，根据 ERP 图像的纬度来调整卷积的接受域。Tateno 等[7]提出了一种可变形卷积滤波器，基于单位球坐标对切平面上的像素网格进行采样。Pintore 等[8]提出了 SliceNet 网络，该网络使用规则卷积过滤器直接处理 ERP。SliceNet 仅沿垂直方向减少输入张量，收集垂直切片序列，并采用 LSTM 网络恢复切片之间的长短期空间关系。还有部分方法利用全景图像的立方体投影（Cubic Projection，CP）无畸变和不失真的特点做为网络输入，如 Wang 等[9]提出的 BiFuse 网络，该网络引入了一个双分支网络结构，其中一个分支处理 ERP 输入，另一个分支提取 CP 特征，并对 CP 投影进行圆形填充，将两个分支的语义和几何信息结合起来。受 BiFuse 的启发，Jiang 等[10]提出了 UniFuse，并设计了一种更有效的融合模块，将两种特征结合起来，该网络存在双分支输入，但是只有一个解码器，仅在解码阶段将 CP 特征单向馈送给 ERP 特征。为更好地提取全局上下文信息，Bai 等[11]将 CP 投影与 Transformer 相结合，提出了 GLPanoDepth 网络，首先将 ERP 输入转换为一组 CP 图像，再通过 ViT 模型建模长距离依赖关系，该网络同样为双分支网络，其中一个为 ERP 卷积，另一个分支为 ViT Transformer，并在解码器中融合两者特征。Rey 等[12]提出了 360MonoDepth，将输入图像投影到一组切面上，产生透视图，为获得全局一致的视差估计，使用可变形的多尺度对齐和梯度域混合来重新组合透视图像深度估计的结果。

还有一些方法通过几何位置信息嵌入来减少畸变，Li 等[13]提出了 OmniFusion 网络，将全景图像转换为畸变较小的透视图像（即切线图像），通过 CNN 获得逐块预测，引入了额外的 3D 几何嵌入，以减轻特征的差异，然后合并逐块结果作为最终输出。Ai 等[14]在其基础上提出了 HRDFuse 网络，通过设计的空间特征对齐模块和协同深度分布分类模块，学习特征相似性与深度整体区域分布特征，最终通过深度直方图中心的线性组合预测深度。Shen 等[15]提出了 PanoFormer 网络，将球面切线域上的 patch 划分为 token，以减少全景畸变的负面影响，该网络设计了一个引入球面几何结构的自注意力模块。Ai 等[16]提出了一种新的框架 Elite360D，输入 ERP 图像和二十面体投影点集，该点集不失真且空间连续。Carlsson 等[17]利用融合 HEALPix 网格[18]和 Swin Transformer 进行鱼眼图像分割和深度估计，但其融合的程度有限，只适用于 2/3 球面，并不能应用到整个球面上。

本文提出一种融合 HEALPix 网格和 Swin Transformer 的全景图像深度估计方法。该模型引入 HEALPix 网格处理全景图像，将全景图像转换为 HEALPix 网格的 Nest 排序格式，直接对全景图像的球面数据进行处理，有效缓解球面几何特性导致的图像失真问题。此外，将 HEALPix 网格的球面分层特性与 Swin Transformer 的自注意力机制相结合，使自注意力的窗口移动和掩码操作能够适配 HEALPix 网格的 Nest 排序格式，使得模型可以通过自注意力机制有效地捕捉长距离依赖关系和局部特征，进一步提升深度估计的效果。

1　HEALPix 网格处理

全景图像具备球面几何特性，可覆盖 360°×180°场景的完整视野，为便于采用常规方法对其进行分析与处理，需先将其映射至平面空间。当前主流的全景图像平面投影方式包括 ERP 投影和 CP 投影。

ERP 投影将球面按照经度和纬度投影到矩形平面。如图 1 所示，球面的南北极区域投影到矩形平面后被拉长，产生局部失真。CP 投影则将球面投影到外切立方体的 6 个面上，每个面对应球面上的

一部分区域。相较于 ERP 投影,避免了南北极地区因为投影导致的失真问题,但是在边界(如图 1(b)的上、下投影处)部分存在不连续性。



(a) ERP 投影 (b) CP 投影

图 1 全景图像投影方式

Fig. 1 Panoramic image projection method

为避免全景图像投影到平面造成局部失真和图像不连续等问题,本文采用 HEALPix 网格对全景图像数据进行处理,来保持全景图像的原始球面结构,HEALPix 是一种分层等面积等纬度处理球面数据的方法,被广泛应用于天文学、地理信息系统、气候建模等领域。HEALPix 网格采用递归嵌套分层的方式对球面进行划分,使球面上每个像素的面积相等。ERP 投影将球面划分为均匀的经度和纬度线,像素越靠近南北极地区,像素密度会越来越大,而赤道区域的像素密度却过于稀疏,但赤道区域的面积却比南北极面积要大,因此在进行球面数据处理时会对南北极地区进行过度的计算。与 ERP 投影不同,HEALPix 网格能够避免这种情况。其独特的球面数据处理方式使得南北极地区和赤道地区的像素密度相同,因此在进行球面数据处理时,尤其是对高分辨率图像,在每个区域的注意力和算力相同,其划分如图 2 所示,图中每一个方块代表着一个像素。

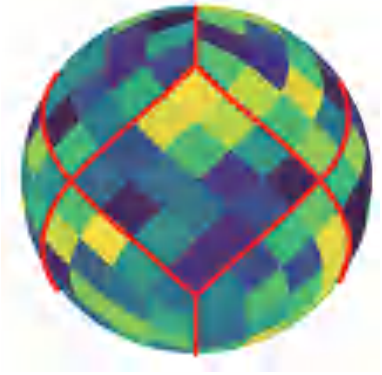


图 2 HEALPix 网格划分

Fig. 2 Division of the HEALPix grid

HEALPix 网格的构建基于一种层次化的递归方法,其处理后的数据有两种排序方式,Nest 和 Ring,本文使用的为 Nest 排序,初始情况下,球面被划分为 12 个区域,图 2 的红色方块代表着一个基本区域,每个区域的面积相同,且均匀分布在球面上。这 12 个区域代表了球面上的基本单元,之后在每个区域内部进一步进行分割,递归细化这些区域,增加网格的分辨率,为保持分割后每个像素的面积相等,每次细分都会将每个四边形像素分成 4 个子四边形,重复这一过程,可以生成任意分辨率的 HEALPix 网格,具体如图 3 所示。

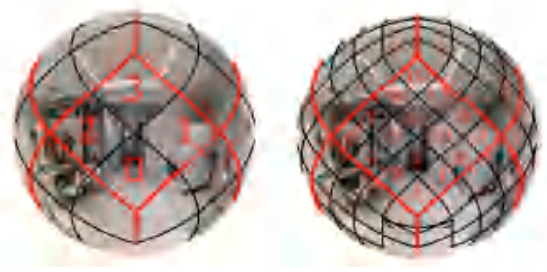


图 3 不同分辨率的 HEALPix 网格

Fig. 3 HEALPix grids of different resolutions

在高分辨率的情况下,像素数量随之增加,从而可以更细致地描述球面数据。HEALPix 可根据不同尺度的需求进行不同分辨率的细分,满足从低分辨率到高分辨率的各种精度需求。随着层次递归的加深,网格分辨率逐步提升,使得 HEALPix 在不同尺度的数据处理中具备很好的灵活性和适应性。分辨率与像素总数相关,计算公式如下:

$$N_{\mathrm{Pix}} = 12N_{\mathrm{Side}}^2 \tag{1}$$

式中:N_{Pix} 为像素总数,N_{Side} 为单个区域边上的像素数,如图 3 所示,示例中 $N_{\mathrm{Side}} = 4$。

2 网络结构

Swin Transformer 在语义分割和深度估计等密集回归任务中表现出色,其核心思想是将输入特征

划分成互不重叠的局部窗口，在局部窗口内进行自注意力计算，并通过窗口移位来增加窗口之间的特征交互，从而实现特征全局注意力的交互，由于自注意力是在局部窗口内计算的，因此相比全局注意力，大幅度减少了因像素数量增加而带来的计算负担，因此，本文采用了编码器解码器结构的 Swin Transformer 变体。该变体引入 U-Net 网络的跳跃连接机制，并扩展了 Swin Transformer 的编码器部分，解码器层的结构与编码器层类似，主要区别在于补丁合并层被补丁扩展层所替代，从而在解码过程中能将特征逐级恢复到与输入特征相同的尺寸，跳跃连接可以优化细节处理。提升深度估计的精度。网络结构如图 4 所示。(a)是 ERP 投影格式的全景图像；(b)是网络的输入，格式为 Nested 排序；(c)是编码器；(d)是解码器；(e)为网络的输出，格式为 Nest 排序；(f)是真实深度值。

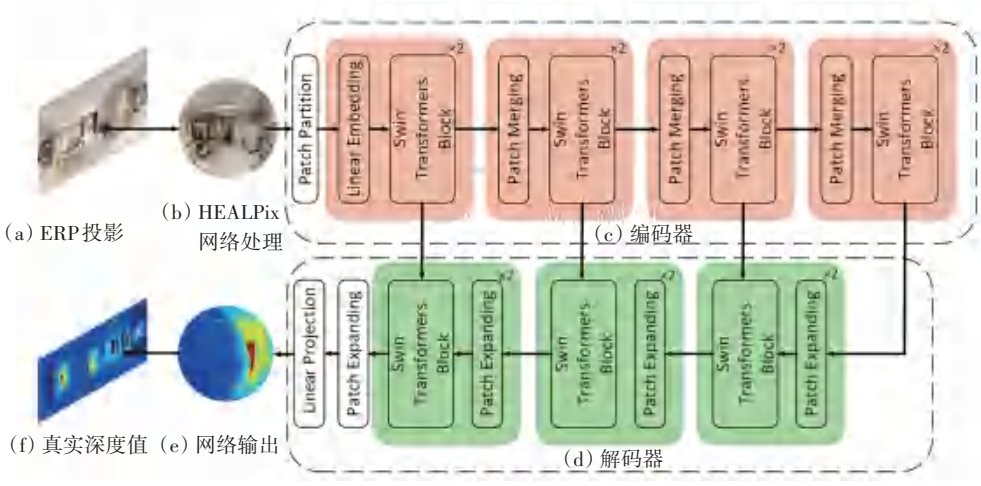


图 4　网络结构

Fig. 4　Network Architecture

2.1　网络主体结构

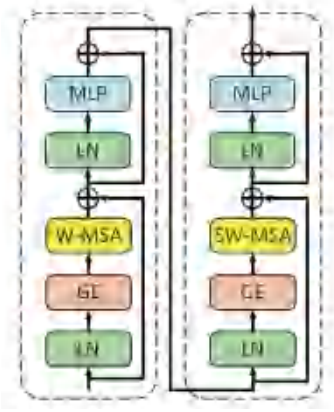
U-net 网络由 Ronneberger 等[19]提出，在医学图像分割取得了良好的效果，后来该网络在深度估计上也取得了较好的效果。本文网络也是一种 U 型对称结构，主要结构包含编码器、解码器和跳跃连接，网络结构如图 4 所示。在本文的网络中，Patch 与原始 Swin Transformer 中的底层矩形像素网格无关，但与球体上的 HEALPix 网格有关，将 Swin Transformer 的 Patch、Window 和 Shift 机制适配 HEALPix 网格，使 Swin Transformer 能够对全景图像的球面数据进行处理。

模型的流程如下，首先将 ERP 格式的全景图像使用 HEALPix 处理成 Nest 排序的一维张量，此时形状为(B, 3, N)，将其作为网络的输入，接着降低初始分辨率，采用 Patch 拆分模块(Patch Partition)将该一维向量拆分成长度为 P 的非重叠补丁(Patch tokens)，通过线性嵌入层(Linear Embedding)将通道倒数提升到 C，得到形状为 (B, C, N/P) 的张量，并进行几何位置嵌入，随后将这些补丁(Patch tokens)输入 Swin Transformer 模块。如图 5 所示，Swin Transformer 模块输入和输出的尺寸形状一致，并在每个窗口(Windows)内执行注意力计算，在第二个 Swin Transformer 模块中，为实现全局信息交互，将窗口沿着 HEALPix 网格轴移动半个窗口的大小，从而实现跨越窗口操作，此时，将其它窗口的补丁(Patch tokens)进行移动来填充当前窗口，并进行自注意力掩码计算。掩码操作不会对移动后的补丁(Patch tokens)执行注意力计算，从而避免了不相邻像素之间的干扰，并使得信息可以在不同窗口之间进行有效地传播。为降低特征的分辨率，补丁会在网络中逐步合并。在解码阶段，补丁合并层(Patch Merging)会被补丁扩展层(Patch Expanding)所取代，补丁扩展层(Patch Expanding)会逐级将特征恢复到与原始输入相同的分辨率，达到最终输出与输入分辨率一致的结果。跳跃连接在解码器部分插入，可以对特征进行细节处理。

2.2　窗口移动

为了在图像中实现全局信息的交互，原始 Swin Transformer 在每两个注意力层中，会进行一次窗口移动操作，沿着图像的两个轴方向将窗口(Window)平移半个窗口的大小来实现不同窗口之间的交互。在本文的网络中，输入特征是 HEALPix 处理后的 Nest 格式一维排序数据，需要将这种窗口移动操作应用于 HEALPix 网格上，使模型能够直接处理球面

全景数据。本文选择在 HEALPix 网格上沿垂直方向进行平移,每次移动半个窗口的大小。与原始 Swin Transformer 的平移方案类似,这种平移操作也是在两个方向上分布完成的,从而得到了平移结果,图 6 展示了这一过程。通过重新排列像素来填充所有窗口,随后使用掩码机制(Masking),避免注意力机制关注来自移动后的补丁(Patch tokens)。



LN:层归一化; GE:几何嵌入卷积; W-MSA:窗口多头自注意力; SW-MSA:移位窗口多头自注意力; MLP:多层感知机

图 5 改进的 Swin Transformer 模块结构

Fig. 5 Structure of the improved Swin Transformer Block

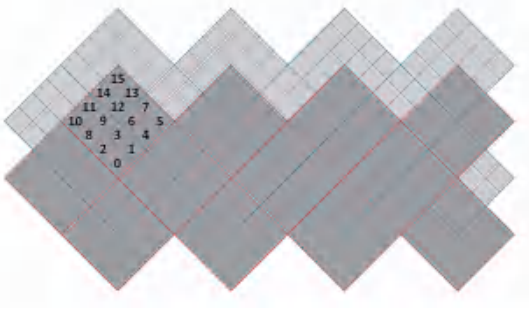


图 6 窗口移动方案

Fig. 6 Window movement scenarios

图中 $N_{\text{side}}=4$,图中的编号代表着像素索引的位置,窗口大小为 16,平移前的窗口用红框标出,补丁按嵌套方案编号。平移半个窗口大小后,补丁被划分到用蓝框框住的窗口中,例如:补丁 0 在平移后变成了补丁 12。

2.3 相对位置偏移

传统的自注意力机制进行自注意力计算时未考虑特征的相对位置,即每个位置对其他位置的注意力权重计算仅依赖于其特征值,对于图像等具有空间结构的数据来说,相对位置非常重要,这种方法会导致空间关系的丧失。因此引入了相对位置偏移,在执行自注意力计算时添加查询(Query)和键(Key)之间的相对位置权重来增强空间感知能力。此时的自注意力计算公式如下:

$$\text{Att}(Q,K,V)=\text{Softmax}(QK^{\text{T}}/\sqrt{d}+\boldsymbol{B})V \tag{2}$$

其中,偏置矩阵 $\boldsymbol{B}$ 是可学习的参数,其仅与补丁(Patch tokens)之间的位置差异有关。所有具有相同位置的补丁(Patch tokens)偏置值一样。在 HEALPix 网格中,每个基四边形内部的特征排列近似一个矩形网格,可利用该矩形网格计算补丁(Patch tokens)之间的相对位置,以便获得相对位置偏置的映射矩阵 $\boldsymbol{B}$,大小为 $[n_{\text{win}}\times n_{\text{win}}]$,计算公式如下:

$$\boldsymbol{B}_{ij}=\hat{\boldsymbol{B}}_{ab} \tag{3}$$

$$a=x(i)-x(j)+\sqrt{n_{\text{win}}} \tag{4}$$

$$b=y(i)-y(j)+\sqrt{n_{\text{win}}} \tag{5}$$

式中 $\hat{\boldsymbol{B}}$ 是可学习矩阵,其包含了关于补丁(Patch tokens)之间的相对偏置值,这些值是通过训练数据学习得到的,相同位置差异的相对偏置值相同。这个矩阵的大小取决于 n_{win},与窗口大小有关。a,b 是像素的位置索引公式,如图 5 中的像素 1,$x(i)=1$, $y(i)=0$。这种相对位置偏置映射的引入,使得在进行自注意力计算时考虑到了空间位置信息,增强了模型对空间结构的感知能力,使得 Swin Transformer 能够更有效地在 HEALPix 网格上处理球面数据。

3 实验与分析

实验在 NVIDIA GeForce GTX3070 GPU,32 核 Intel(R) Core(TM) i7-12600KF CPU @ 1.80 GHz 的服务器上进行;软件部分使用 Pycharm 作为集成开发环境(IDE),通过 Python 编程语言以及深度学习 Pytorch 框架进行程序设计。模型训练的优化器选择 Adam,初始化学习率为 0.000 01,设置训练轮次 Epoch 为 100,Batch size 为 2,损失函数为 berHu 损失,公式如下:

$$\text{BerHuLoss}(y,y_i^{\text{t}})=\begin{cases}|y-y_i^{\text{t}}|, & \text{if } |y-y_i^{\text{t}}|\leqslant\sigma\\ \dfrac{(y-y_i^{\text{t}})+\sigma^2}{2\sigma}, & \text{otherwise}\end{cases} \tag{6}$$

式中:y 为预测值,y_i^{t} 为真实值,σ 是一个超参数,本文设置的值为 0.1。

3.1 评价指标

本文采用的评价指标有均值绝对误差(Mean Absolute Error, MAE)、绝对相对误差(Absolute Relative Error, AbsRel)、均方根误差(Root Mean

Squared Error，RMSE）、均方根对数误差（Root Mean Squared Logarithmic Error，RMSLE），以及不同阈值下的深度估计准确率（Acc@δ），各指标公式定义如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-y_i^{\mathrm{t}}\right| \tag{7}$$

$$\mathrm{AbsRel}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i-y_i^{\mathrm{t}}\right|}{y_i} \tag{8}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-y_i^{\mathrm{t}}\right)^2} \tag{9}$$

$$\mathrm{RMSLE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(1+y_i)-\log(1+y_i^{\mathrm{t}})\right)^2} \tag{10}$$

$$\mathrm{Acc}@\delta=\frac{1}{n}\sum_{i}^{n}\prod\left(\max\left(\frac{y_i^{\mathrm{t}}}{y_i},\frac{y_i}{y_i^{\mathrm{t}}}\right)\leqslant\delta\right) \tag{11}$$

其中，n 为像素总数；y_i 为该像素的预测值；y_i^{t} 为该像素的真实值；δ 为超参数，分别取 $\delta_1=1.25$，$\delta_2=1.25^2$，$\delta_3=1.25^3$；$\prod(\cdot)$ 为指示函数，成立时为 1，不成立时为 0。

3.2　对比分析

为验证本文方法在全景图像深度估计方面的有效性，实验在 Stanford2D3D 数据集[20]上与其他方法进行对比。Stanford2D3D 数据集是一个专为深度估计、语义分割和三维重建等任务设计的全景图像数据集。该数据集来自斯坦福大学校园内 6 个不同建筑的真实场景，包含丰富的全景图像、语义和纹理信息。数据集中提供了 RGB 全景图像、深度图、语义分割标注以及相应的 3D 点云数据。这些全景图像通过球面投影捕获，覆盖完整的 360°视野，分辨率为 2 048×1 024，适合全景视觉任务。此外，每个场景还附带了精准的相机参数和几何结构信息，使其成为研究室内三维环境的理想基准数据集。其中对数据集的处理与 UniFuse 一致，将图像的分辨率转换为 1 024×512，并进行翻转、增强处理，实验结果见表 1、图 7。

表 1　不同方法的评价结果

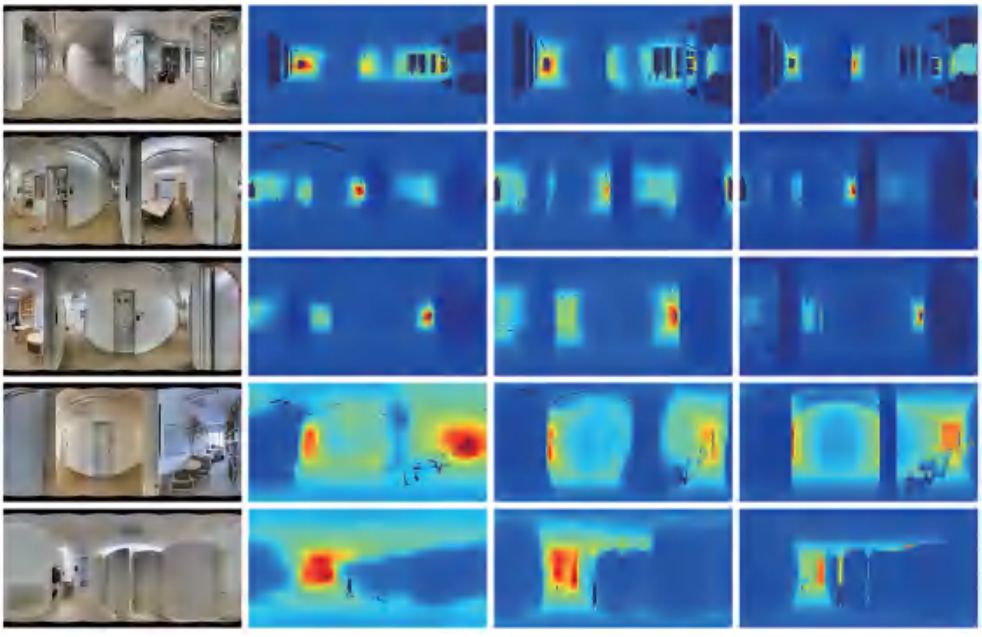
Table 1　Evaluation results of different methods

方法	MAE ↓	AbsRel ↓	RMSE ↓	RMSLE ↓	δ_1 ↑/%	δ_2 ↑/%	δ_3 ↑/%
BiFuse	0.234 3	0.120 9	0.414 2	0.078 70	86.60	95.80	98.60
UniFuse	0.208 2	0.111 4	0.369 1	0.072 10	87.11	96.64	98.82
OmniDepth	0.374 3	0.199 6	0.615 2	0.121 20	68.77	88.91	95.78
SliceNet	0.248 4	0.124 9	0.437 0	0.087 30	83.77	94.14	97.77
RectNet	—	0.199 6	0.615 2	—	68.77	88.91	95.78
HoHoNet	0.202 7	0.101 4	0.383 4	0.066 80	90.54	96.93	98.86
OmniFusion	—	0.095 0	0.347 4	0.159 90	89.88	97.96	99.24
BiFuse++	0.217 3	0.111 7	0.372 0	0.072 70	87.83	96.49	98.84
GLPanoDepth	0.193 2	—	0.349 3	0.068 01	90.15	97.39	99.01
SphereDepth	0.232 3	0.115 8	0.451 2	0.075 40	86.66	96.42	98.63
HRDFuse	—	0.093 5	0.310 6	0.142 20	91.40	97.98	99.27
ACDNet	0.187 9	—	0.341 0	0.066 40	88.72	97.04	98.95
MODE	0.173 6	0.092 7	0.309 6	**0.040 00**	91.69	98.16	99.29
CRF360D	—	0.084 5	0.293 2	—	**92.53**	**98.38**	**99.44**
Elite360D	—	0.118 2	0.375 6	—	88.72	96.84	98.92
本文方法	**0.142 9**	**0.047 2**	**0.097 2**	0.098 50	79.50	93.71	97.61

注：“—”表示该论文并没有使用该评价指标，效果最好的用粗体标出。

本文选择经典的双分支网络 BiFuse 和 UniFuseOmniDepth，以及近年提出的 MODE、CRF360D 和 Elite360D 等网络模型进行对比实验，从实验结果来看，本文方法在多项深度估计评价指标上均取得了显著提升。在 MAE 评估指标上，本文方法的误差为 0.142 9，降低了 17%，在 AbsRel 评估指标上，本文方法为 0.047 2，降低了 44%，在 RMSE 评估指标上，本文方法为 0.097 2，降低了 66%，这表明，本文方法

所提出的融合 HEALPix 网格与 Swin Transformer 网络能够更准确地处理全景图像的球面数据,并在减少预测误差方面优于现有方法,体现了本文方法在准确性上的显著优势,但在 RMSLE 评估指标上相对偏高,说明本文网络在处理近距离物体即深度值偏低的情况时误差偏大。此外,为进一步验证本文方法的深度估计准确率,在阈值分别取 1.25、1.25^2 和 1.25^3 下计算了预测深度值的准确率。尽管本文方法在 MAE、AbsRel 和 RMSE 上取得了显著提升,但准确率表现略逊色于一些对比方法。结合图 7 的测试集样例图可见,本文方法在局部细节方面仍有不足,边界区域区分度不够明显,由于模型本身并没有添加其他模块,后续仍然存在一定的改进空间,可以添加其他模块和分支来提升准确率。



(a) ERP 图像 (b) Swin Transformer (c) 本文 (d) 真实深度值

图 7 不同方法的测试集样例图

Fig. 7 Sample images from the test set of different methods

3.3 消融实验

本节通过消融实验评估所提出的融合 HEALPix 网格与 Swin Transformer 融合方案的有效性。保持其他训练条件相同,将融合网络模型替换成 Swin Transformer 模型,来验证该网络的优越性,实现结果见表 2。由表 2 可见,本文提出的融合网络在各个指标上均优于 Swin Transformer 网络,验证了本文网络提出融合 HEALPix 网格和 Swin Transformer 融合结构的有效性。

表 2 与 Swin Transformer 对比实验

Table 2 Comparative experiments with Swin Transformer

方法	MAE ↓	AbsRel ↓	RMSE ↓	RMSLE ↓	δ_1 ↑/%	δ_2 ↑/%	δ_3 ↑/%
Swin Transformer	0.171 7	0.047 6	**0.091 5**	0.110 1	73.77	91.61	96.97
本文方法	**0.142 9**	**0.047 2**	0.097 2	**0.098 5**	**79.50**	**93.71**	**97.61**

4 结束语

为解决全景图像在平面投影的过程中,图像出现局部失真和不连续等问题,本文提出了一种融合 HEALPix 网格和 Swin Transformer 的深度估计模型,该模型通过 HEALPix 网格将全景图像变成 HEALPix 网格的 Nest 排序,直接对全景图像的球面数据进行处理,从而避免全景图像的失真和不连续问题。通过将 HEALPix 网格的球面分层分步方式与 Swin Transformer 的自注意力机制相结合,使得自注意力的窗口移动和掩码操作能适用于 HEALPix 网格的 Nest 排序,达到能够捕捉全景图像的全局深度信息。在对比实验中,本文方法在多个深度评估标准上均取得了显著的性能提升,尤其是在 MAE

(0.142 9)、AbsRel(0.047 2)和 RMSE(0.097 27)等指标上均优于现有的基于深度学习的全景图像深度估计方法。这表明本文提出的直接对全景图像进行球面处理的方法能够有效降低深度估计中的误差，然而，尽管在误差指标上表现优异，本文方法在准确率上略显不足，未来工作可通过引入其他模块或分支结构，进一步优化模型准确率。

参考文献

[1] 朱哲，黄莉，王宗阳. 融合感知卷积和注意力机制的全景图深度估计方法[J]. 计算机技术与发展，2024，34(12)：16-24.

[2] 贺思. 基于神经网络的全景图像深度估计与场景重建[D]. 北京：北京邮电大学，2022.

[3] 周晓吉. 基于多尺度特征融合的单目深度估计算法[J]. 智能计算机与应用，2024，14(9)：34-40.

[4] 方俊，刘红喜，穆晓飞. 卷积神经网络与 Transformer 混合的自监督单目深度估计算法[J]. 智能计算机与应用，2024，14(1)：168-174.

[5] Sun C，Sun M，Chen H T. Hohonet：360 indoor holistic understanding with latent horizontal features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2021：2573-2582.

[6] Zioulis N，Karakottas A，Zarpalas D，et al. Omnidepth：Dense depth estimation for indoors spherical panoramas[C]//Proceedings of the European Conference on Computer Vision. Cham：Springer，2018：448-465.

[7] Tateno K，Navab N，Tombar F. Distortion-aware convolutional filters for dense prediction in panoramic images[C]//Proceedings of the European Conference on Computer Vision. Cham：Springer，2018：707-722.

[8] Pintore G，Agus M，Almansa E，et al. Slicenet：Deep dense depth estimation from a single indoor panorama using a slice-based representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：11536-11545.

[9] Wang F E，Yeh Y H，Sun M，et al. Bifuse：Monocular 360 depth estimation via bi-projection fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2020：462-471.

[10] Jiang H，Sheng Z，Zhu S，et al. Unifuse：Unidirectional fusion for 360 panorama depth estimation[J]. IEEE Robotics and Automation Letters，2021，6(2)：1519-1526.

[11] Bai J，Qin H，Lai S，et al. Glpanodepth：Global-to-local panoramic depth estimation[J]. IEEE Transactions on Image Processing，2024，33：2936-2949.

[12] Rey-Area M，Yuan M，Richardt C. 360 monodepth：High-resolution 360° monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2022：3762-3772.

[13] Li Y，Guo Y，Yan Z，et al. Omnifusion：360 monocular depth estimation via geometry-aware fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2022：2801-2810.

[14] Ai H，Cao Z，Cao Y P，et al. Hrdfuse：Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2023：13273-13282.

[15] Shen Z，Lin C，Liao K，et al. PanoFormer：Panorama transformer for indoor 360° depth estimation[C]//Proceedings of European Conference on Computer Vision. Cham：Springer，2022：195-211.

[16] Ai H，Wang L. Elite360D：Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and PatternRecognition. Piscataway，NJ：IEEE，2024：9926-9935.

[17] Carlsson O，Gerken J E，Linander H，et al. Heal-swin：A vision transformer on the sphere[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2024：6067-6077.

[18] Gorski K M，Hivon E，Banday A J，et al. HEALPix：A framework for high-resolution discretization and fast analysis of data distributed on the sphere[J]. The Astrophysical Journal，2005，622(2)：759.

[19] Ronneberger O，Fischer P，Brox T. U-net：Convolutional networks for biomedical image segmentation[C]//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham：Springer，2015：234-241.

[20] Armeni I，Sax A，Zamir A R，et al. Joint 2D-3D-semantic data for indoor scene understanding[J]. arXiv preprint arXiv，1702.08765，2017.