arXiv:2301.05845v1 [cs.CV] 14 Jan 2023

S^2Net：球面上的精准全景深度估计

Meng Li[1], Senbo Wang[1], Weihao Yuan[1],Weichao Shen[1],Zhe Sheng[1] 和 Zilong Dong[1]

摘要——单目深度估计是一个具有歧义的问题，因此，全局结构线索在当前数据驱动的单视图深度估计方法中发挥着重要作用。全景图像通过等距柱状投影捕获周围环境的完整空间信息，但这种投影会引入严重畸变。因此，深度估计方法需要能够处理畸变，并从图像中提取全局上下文信息。本文提出了一种用于在单位球面上进行单目全景深度估计的端到端深度网络。具体而言，我们将从等距柱状投影图像中提取的特征图投影到由均匀分布的网格采样得到的单位球面上，使解码器网络能够聚合畸变减小后的特征图中的信息。同时，我们提出了一种基于全局交叉注意力的融合模块，用于融合来自跳跃连接的特征图，并增强获取全局上下文的能力。我们在五个全景深度估计数据集上进行了实验，结果表明，所提方法显著优于以往最先进的方法。所有相关代码将在未来几天内开源。

索引词—**视觉感知深度学习、全向视觉、深度学习方法**

I. 引言

PREDICTING 仅从单张图像中估计深度尤其具有挑战性，因为这个问题本质上存在歧义 [1]。当前的数据驱动方法通过学习隐式结构和语义信息来预测深度，这些信息为深度估计提供了充足的单目线索。这要求网络观察场景中更大的区域 [2],，近年来，单视图深度估计领域的许多研究已对此进行了探索 [3], [4]。

以往的工作 [3] 通过在编码器和解码器之间添加非局部深度注意力模块，来解决扩展感受野的问题，但训练该模块需要额外监督。Adabins [4]使用由视觉Transformer组成的后处理模块，来优化传统编码器-解码器架构的输出，但网络中“传统”部分的感受野并未扩大。Ranftl等人 [5] 将视觉Transformer用作特征提取模块，使特征图在每个阶段都具有全局感受野，与全卷积网络 [5]相比，取得了显著改进。然而，由于使用传统卷积网络作为解码器，其编码非局部几何依赖的能力仍然有限。全景深度估计方法 SliceNet [6]使用循环神经网络来增强其检索全局几何信息的能力，但它要求全景图像满足重力对齐这一额外约束。

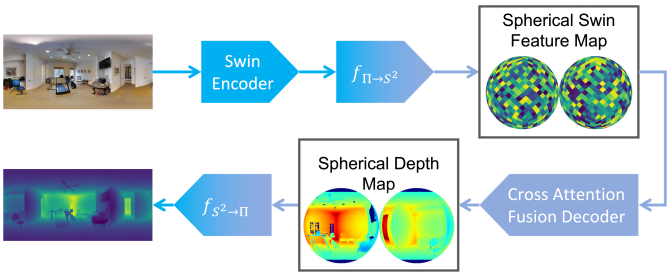

图1：所提方法的流程。浅蓝色部分作用于平面ERP图像，而灰蓝色部分作用于单位球面。$\mathbf{f}_{\Pi \to \mathbf{S^2}}$ 表示从图像平面到单位球面的映射，而$\mathbf{f}_{\mathbf{S^2} \to \Pi}$ 表示从单位球面到图像平面的映射。

除了扩展感受野方面的挑战外，对于全景图像，其独特的**等距柱状投影（ERP）**图像表示也带来了额外挑战。这会造成严重畸变，直接处理ERP图像的网络会因这些畸变而产生偏差，使其偏向极点 [7], ，因为ERP图像中的像素在对应球面上的分布并不均匀。此外，直接从ERP图像表示中学习，无法满足全景图像的隐含假设：图像左侧与右侧在物理上相连，从而导致深度图 [8]的空间一致性问题。

一些先前方法 [9], [10]采用额外分支，引入由全景图像生成的立方体贴图，以补偿畸变。然而，立方体贴图不同面的接缝处会出现不连续性，因此无法完全补偿ERP图像中的畸变。其他方法 [11], [12], [13] 则采用不规则卷积核或空洞卷积，但其去畸变能力受网络学习能力的限制。此外，上述所有方法都需要循环填充来确保深度图的空间一致性，但这一问题仍无法得到彻底解决。

总而言之，这些问题要求全景深度估计方法能够以一种无畸变的方式表示 ERP 图像，

[1]阿里巴巴集团，中国杭州市登彩街1008号，邮编310000 wsb_pro@live.com, shengzhe.sz@alibaba-inc.com, list.dzl@alibaba-inc.com Meng Li 和 Senbo Wang 对本文贡献相同。数字对象唯一标识符 10.1109/LRA.2023.3234820 c◯ 2023 IEEE。允许个人使用本材料。任何其他用途均须获得 IEEE 许可，无论是在当前还是未来的任何媒体上，包括为广告或促销目的而重印/再版本材料、创作新的合集作品、转售或再分发至服务器或列表，或在其他作品中重复使用本作品中任何受版权保护的部分。

在这种表示方式中，像素均匀分布，并且拥有更大的感受野，从而能够更好地探索其中蕴含的全局结构信息。

受先前球面卷积神经网络相关工作的启发， [14], [15], 我们使用 HEALPix 采样方法，将全景深度估计任务转移到由 ERP 图像生成的球面上。 [16],该方法可确保采样像素在球体上均匀分布，从而大幅减少失真。此外，球面表示天然保证了全景图像两侧的空间一致性，因此减少了对循环填充等补救方法的需求。

我们还引入了**交叉注意力融合（CAF）**模块，该模块利用可学习的交叉注意力融合来自跳跃连接的特征图，增强解码器获取全局上下文信息的能力。所提方法的流程见图1。据我们所知，所提方法是首个在真实世界视觉任务中于单位球面上应用视觉Transformer的方法。

所提方法在几乎所有公开数据集上的表现都显著优于先前方法，尤其是在侧重评估方法泛化能力的新 Pano3D [7] 数据集上： $RMSE$ 误差降低了**23.3%**, $AbsRel$误差降低了**16.3%**。我们的方法在空间一致性以及对重力方向错位的鲁棒性方面也取得了良好效果。

本文的主要贡献如下：

1）提出了一种新颖的端到端深度网络，将解码器网络移至球面上，从而消除 ERP 图像的畸变，并保持深度图的空间一致性。2）提出了一种基于球面 Transformer 的融合模块（CAF），以可学习的方式融合来自跳跃连接的特征，并增强获取全局上下文信息的能力。

本文的组织结构如下：第二节回顾并讨论相关工作；第三节介绍所提方法的必要细节；第四节展示实验和消融研究；第五节总结全文。

II. 相关工作

a) 透视相机图像的深度估计：新方法采用了知识蒸馏 [17], 、基于注意力的方法 [4], [18] 和自监督 [19] 等思路。拓展网络的全局上下文感知能力也是一个重点研究方向。方法 [4], [3] 尝试在网络中加入基于视觉 Transformer的处理模块，以拓展方法的全局上下文，但前者工作中网络的传统部分并未扩大其全局上下文，后者工作 [3]则需要对模块进行额外监督。先前的工作 [5]将基于注意力的模块作为特征提取模块，但其解码器仍采用传统结构。

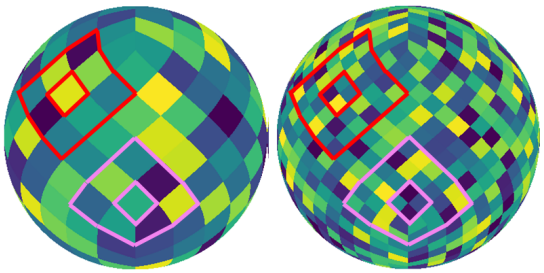

图 2：我们通过两个具有不同 n_{size} 的示例展示 HEALPix 的多尺度采样能力：左侧为 $n_{size} = 4$ ，右侧为 $n_{size} = 8$ 。

b) 全景图像的深度估计：一些方法 [9], [10] 使用立方体贴图作为图像表示方法，并尝试利用立方体贴图中的信息来补偿 ERP 图像的畸变，但结果可能会受到立方体贴图各面边缘不连续性的影响。另一些方法 [11], [12]则采用空洞卷积等专门的卷积方法来处理 ERP 图像中的畸变，并扩大网络的感受区域。这些方法没有采用减少畸变的表示方法来表示 ERP 图像，因此结果可能会受到影响。一些方法尝试使用一维表示 [6], [20]来表示 ERP 图像；然而，一维表示可能无法完全消除纬度方向上的畸变。与这些先前工作相比，所提方法采用球体上的减畸变表示，不受立方体贴图不连续性的影响，并且可通过视觉方法进一步扩大所提方法的感受野。

c）球面深度学习：这些方法大多在全景图像理解任务中使用空洞卷积或特殊形式的卷积核 [21], [22], ，或者利用纬度和经度 [23], [24] 补偿畸变。与所提方法相似的工作有两项：一项工作 [25] 采用二十面体采样方法 [26], 并构建了一种用于 ERP 图像表示学习的新型 Transformer。另一项工作 [14] 采用 HEALPix [16]作为采样方法，并针对这种独特的像素表示提出了一种卷积神经网络方法。与上述方法相比，我们的方法是首个在单位球面上完整实现视觉变换技术的方案，并显著提高了全景深度估计的准确率。

III. 方法

我们采用经典的带有跳跃连接的编码器-解码器深度网络架构。视觉Transformer编码器网络从输入 ERP 图像中提取多尺度特征图，并将其从平面表示转换为均匀球面像素表示。随后，多尺度解码器网络沿球体整个表面执行深度回归，以预测深度估计结果。该多尺度解码器网络由单位球面上不同尺度的若干融合模块构成，用于融合

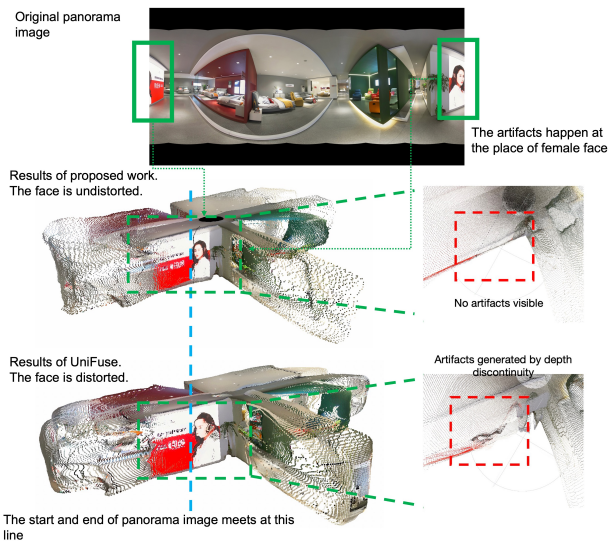

图 3：我们比较了 UniFuse（**采用循环填充**） [10] 与所提方法的空间一致性。UniFuse 等方法无法保证 ERP 图像左右两端的空间一致性，而所提方法解决了这一问题。

来自跳跃连接的球面特征图，其中包含在编码器部分学习到的细粒度局部细节。

A. 全景图像表示

我们使用 HEALPix（分层等面积等纬度像素化） [16] 在单位球面上生成均匀分布的多尺度像素；每层采样 n_{size} 个样本时，采样像素总数为 $n_{pix} = 12 * {n_{size}}^2$。我们使用由粗到细的 4 个球面层级，其像素数量分别为 $[192, 768, 3072, 12288]$。图 2 展示了球面网格示例。

值得注意的是，在 HEALPix 中使用较少（下采样）或较多（上采样）像素都不会对最终结果产生积极影响。相关测试见表 I。我们分别将像素数量上采样和下采样至 $[768, 3072, 12288, 12288]$ 和 $[48, 192, 768, 3072]$，以研究其对最终准确率的影响。由于 GPU 内存有限，我们不对最精细层级进行上采样。所提像素数量取得了整体最佳的准确率，原因可能有两方面：像素较少时，单个像素覆盖的区域较大，所携带的全局上下文信息更多，但细节较少；像素较多时，网络能够在深度图中获取更多细节，但会丢失全局上下文信息。所提像素数量似乎在这两个因素之间取得了良好的平衡。

HEALPix 有助于定义球面与图像平面之间的两种映射：从图像平面到单位球面的映射 $\mathbf{f}_{\Pi \to \mathbf{S}^2}$，以及从单位球面到图像平面的映射$\mathbf{f}_{\mathbf{S}^2 \to \Pi}$。这两种映射可使用开源软件[1]预先计算；预先计算的映射可用于为每个像素构建邻域关系查找表，从而

[1]https://healpy.readthedocs.io/en/latest/

表 I：不同 HEALPix 像素数量对 M3D数据集上最终深度图准确率的影响比较 [27]。

像素数量	误差 ⇓ $AbsRel$	$RMSE$	$RMSLE$	准确率（%） ⇑ $\delta < 1.25$
下采样	0.0902	0.4211	0.0582	92.07
原始	**0.0865**	**0.4052**	**0.0559**	**92.63**
上采样	0.0866	0.4109	0.0565	92.54

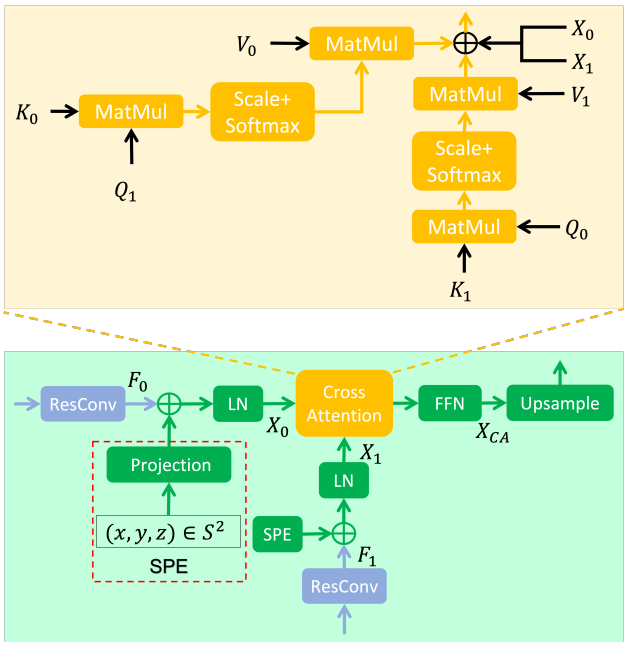

图 4：在球体上进行自注意力计算的 CAF模块结构。“LN”、“SPE”和“ResCon”分别表示层归一化、球面位置嵌入（第 III-C 节）和残差卷积。

确保能够轻松实现自注意力计算。我们使用双线性插值计算 HEALPix 球面上每个像素的颜色；使用其他插值方法对最终准确率几乎没有影响。编码器处理原始 ERP 图像，并使用 HEALPix 将生成的不同通道数的特征图映射到球面上，使通道数统一为 $D \in \{256, 320, 512\}$。基于单位球面的图像表示没有边界，因此单位球面上的操作天然具有循环性。与 UniFuse [10] 等采用循环填充的方法相比，所提方法更能维持 ERP 图像左右两侧的空间一致性，因为感受野和填充宽度都有限。图 3 展示了一个直观的示例。

B. 交叉注意力融合

多个作用于多尺度球面特征的**交叉注意力融合（CAF）**模块共同构成多尺度解码器。CAF旨在增强网络获取全局上下文的能力，并融合来自跳跃连接的特征图，以增强估计深度图的细节。

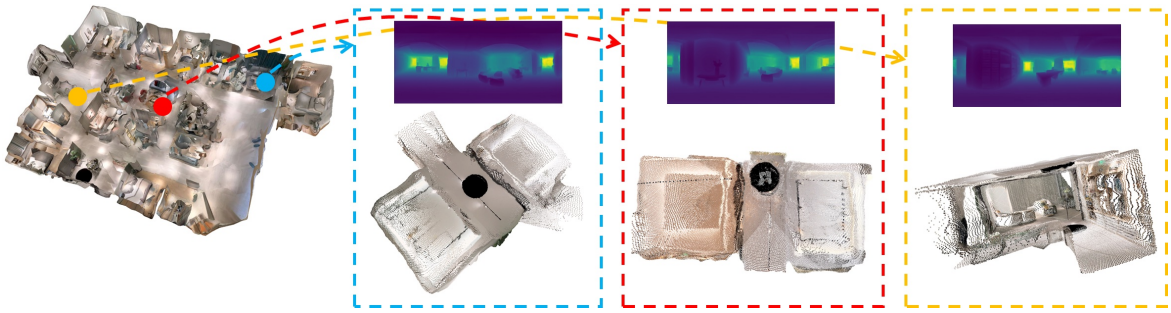

图5：我们在真实世界场景的不同位置采集了一系列全景图像，并为每幅图像估计一张深度图。随后，通过运动恢复结构将这些深度图配准到同一全局坐标系中。图左侧展示了根据配准结果生成的最终网格模型，右侧展示了场景中三个示例位置的深度图及其对应的点云。

表II：比较SPE中坐标表示方式对M3D数据集 [27]上最终深度图准确率的影响。我们用 $[lat, lon]$ 表示使用点的经度和纬度信息的SPE。

坐标	误差 ⇓ $AbsRel$	$RMSE$	$RMSLE$	准确率（%） ⇑ $\delta < 1.25$
$[x, y]$	0.0910	0.4238	0.0589	91.83
$[y, z]$	0.0900	0.4202	0.0581	92.21
$[x, z]$	0.0874	0.4162	0.0570	92.60
$[lat, lon]$	0.0954	0.4416	0.0616	90.92
$[x, y, z]$	**0.0865**	**0.4052**	**0.0559**	**92.63**

CAF采用“全局”交叉注意力计算机制，即在全局范围内计算不同位置特征图像素之间的相关性。通过这种方式，CAF具备建模长程依赖的能力，并能够从多路径特征图中学习更丰富的全局上下文信息。跳跃连接中的特征来自金字塔编码器网络，因此包含“局部”的细粒度细节；而来自解码器的特征则相对“全局”，具有更大的感受野。以往的全局-局部融合策略大多采用简单的求和或拼接操作，直接将这些具有不同上下文的特征图混合在一起。我们的CAF模块采用基于可学习全局注意力的融合策略：将具有局部上下文的跳跃连接特征图视为“引导”，用于增强具有全局上下文的解码器特征图，反之亦然。通过这种方式，CAF通过学习补偿值来增强两种特征图之间的相关性。CAF模块的详细结构见图4。

我们使用基础表示采用直接相加的融合模块，使用融合表示采用我们CAF的融合模块。基础融合模块仅由文献 [5] 中的两个简单残差卷积单元组成，采用以往工作 [14], [25]中的基础球面卷积方法。

给定来自跳跃连接和解码器网络的特征图 F_i 以及位置嵌入 ζ_i，CAF的输入可定义为：

$$X_i = LN(F_i + \zeta_i), i \in \{0, 1\} \tag{1}$$

CAF模块 le 尝试估计两个补偿项 Att_0

表III：位置嵌入方法对M3D数据集上最终深度图准确率的影响比较 [27]

编码方法	误差 ⇓ $AbsRel$	$RMSE$	$RMSLE$	准确率（%） ⇑ $\delta < 1.25$
可学习	0.0881	0.4142	0.0570	92.369
SPE	**0.0865**	**0.4052**	**0.0559**	**92.634**

表 IV：不同损失函数对 M3D 数据集上最终深度图准确率的影响比较 [27]。我们使用“RMSERel”表示相对均方根误差 [29]。

损失函数	误差 ⇓ $AbsRel$	$RMSE$	$RMSLE$	准确率（%） ⇑ $\delta < 1.25$
RMSERel	0.0943	0.4200	0.0597	91.49
均方根对数误差	**0.0865**	**0.4052**	**0.0559**	**92.63**

并通过交叉注意力得到 Att_1：

$$\begin{aligned} Att_0(Q_1, K_0, V_0) &= SoftMax(Q_1 K_0/\sqrt{d})V_0 \\ Att_1(Q_0, K_1, V_1) &= SoftMax(Q_0 K_1/\sqrt{d})V_1 \end{aligned} \tag{2}$$

这两个值被加到输入中：

$$X_{CA} = FFN(LN(X_0 + Att_0 + X_1 + Att_1)) \tag{3}$$

FFN 表示前馈操作， LN 表示层归一化操作。 [25], 这两个操作构成一种残差结构。公式 2 中使用的 Q_i、 K_i 和 V_i 是通过相应的可学习参数 W，从输入 X_0 和 X_1 中生成的：

$$Q_i = W^Q X_i, K_i = W^K X_i, V_i = W^V X_i, i \in \{0, 1\} \tag{4}$$

由于全局注意力的计算需要大量计算资源，我们采用了先前工作 [28] 中使用的方法，通过非重叠窗口划分执行全局子采样交叉注意力。这些划分可以方便地从 HEALPix 的由粗到细结构中推导出来。

C. 球面位置嵌入

为了使 CAF模块更好地获取像素位置信息，我们提出了一种手工设计的位置嵌入方法，即**球面位置嵌入**（**SPE**）。我们沿着

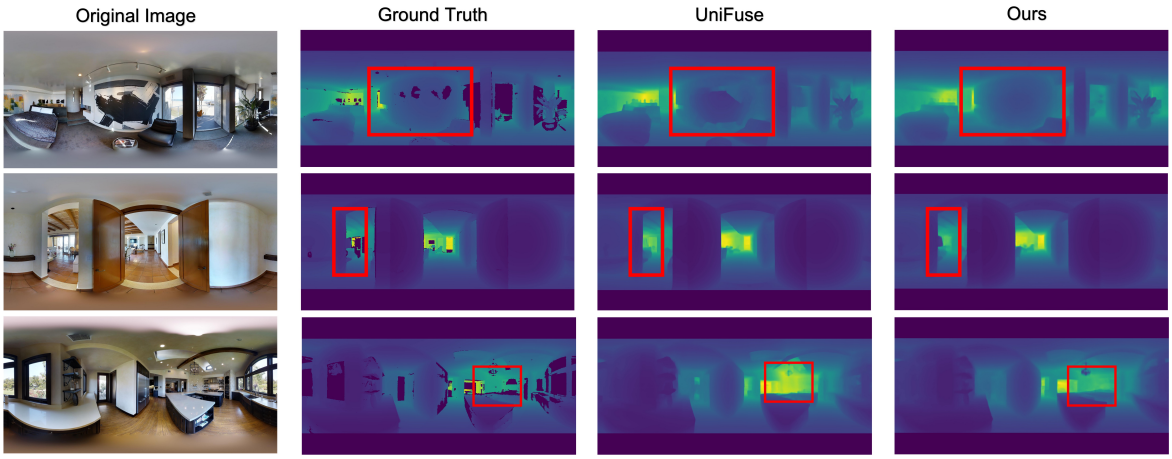

图 6：所提出的方法与 UniFuse [10] 在 M3D数据集中的若干精选场景上的对比。我们用方框标出值得注意的位置。请注意，我们使用不同的颜色来区分不同的显著位置。

通道轴拼接球面坐标，然后将三通道输入嵌入到特征图的维度中：

$$\zeta = C \times [x_{sphere}, y_{sphere}, z_{sphere}]^T \tag{5}$$

其中， $C \in \mathbb{R}^{M\times 3}$ 是一个嵌入矩阵， M 是特征图的通道数。

减少所用坐标的维度，或使用纬度和经度构造位置嵌入，都会降低所提出方法的总体准确率；相关测试见表 II。使用纬度和经度时准确率大幅下降，可能是因为同一片区域的像素位于不同纬度时，经度变化各不相同，这可能会给最终结果引入隐式畸变。

与可学习的位置嵌入机制相比，这种嵌入方法利用球面上的真实空间位置，有助于解码器网络和 CAF 模块考虑嵌入球体中的非欧几里得空间信息，从而获得更好的结果。表 III 展示了这两种嵌入方法的对比结果，所提出的 SPE方法表现更好， $RMSE$ 降低了 **2.2%**。

D. 训练损失

我们使用均方根对数误差（RMSLE）作为损失函数来监督训练过程，如公式 6 所示。 d_i 和d_i 分别表示像素 i 处的估计深度和真值深度。

$$\mathcal{L} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(d_i) - \log(\overline{d_i}))^2} \tag{6}$$

我们针对所有深度值有效的球面像素优化网络参数。均方根对数误差可以减小预测值与真值之间的较大差异对总体损失的影响，并确保损失在不同深度范围内均匀分布。由于球面像素已经在球体上均匀分布，因此每个像素对总体损失的贡献在空间上也是均匀的。我们还注意到，

相对均方根误差 (RMSERel) [29] 也可以减小较大误差的影响；我们在表 IV 中比较了两种损失的准确率结果。RMSERel 的结果不如我们提出的均方根对数误差，这可能是因为均方根对数误差比 RMSERel 更能抑制远处物体带来的损失，因此在全景深度估计任务中能获得更好的空间测量结果，因为这类任务大多关注室内即近距离场景。

IV. 实验

A. 数据集

Matterport3D (M3D) [27] 和 **Stanford2D3D(S2D3D)** [30] 均为使用 Matterport 3D Camera[2],采集的真实世界数据集，而**PanoSUNCG (PSunCG)** [32]和**3D60** [12]均为合成数据集。在上述四个数据集上评估的结果采用了以往研究中的指标[10], [6], [20], ，包括平均绝对误差（MAE）、均方根误差（$RMSE$）、对数均方根误差（$RMSLE$）、绝对相对误差（$AbsRel$）以及三个阈值百分比 $\delta < \alpha^n$ （$\alpha = 1.25, n = 1,2,3$）。

Pano3D [7]是一个用于全景深度估计的评测基准，能够评估网络的泛化能力。我们在 1024×512分辨率下进行了评估，并与 Pano3D[3]提供的基线方法 UNet^{vnl} 和 ResNet$^{comp}_{skip}$ 进行了比较。**请注意，前缀为 w 的指标来自 Pano3D 数据集上的评估，已考虑误差均匀性问题。**

图 5 展示了由所提方法估计的多个深度图生成的对齐网格。图 6 展示了 M3D 数据集中选定场景的 UniFuse [10] 与所提方法生成的深度图对比，结果表明，所提方法能够生成细节更丰富的深度图，并且不易受到具有特殊纹理的图像区域的影响。

B. 实现细节

采用在 ImageNet-22k 上预训练的 Swin-B [33] 作为编码器网络。我们使用相同的数据增强

[2]https://matterport.com/cameras [3] https://github.com/VCL3D/Pano3D

表 V：我们的方法与当前可用方法在 M3D 数据集 [27] 和 S2D3D数据集 [30]上的比较。我们使用 * 标记按中值深度或最大深度进行尺度对齐后评估的结果。 △ 表示未针对全景图像中的畸变进行优化的方法。请注意，本表中 SliceNet [6]的误差和准确率结果取自 ACDNet 的论文 [11] 。

数据集	方法	误差指标 ⇓				准确率指标（%） ⇑		
		MAE	$AbsRel$	$RMSE$	$RMSLE$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
M3D	FCRN* [31]	0.4008	0.2409	0.6704	0.1244	77.03	91.74	96.17
	SliceNet* [6]	0.3296	0.1764	0.6133	0.1045	87.16	94.83	97.16
	我们的方法（基础版）*	0.2025	0.0776	0.3787	0.0513	93.58	98.31	99.44
	我们的方法（融合版）*	**0.1932**	**0.0730**	**0.3720**	**0.0498**	**93.96**	**98.38**	**99.46**
	Bifuse [9]	0.3470	0.2048	0.6259	0.1134	84.52	93.19	96.32
	HoHoNet [20]	0.2862	0.1488	0.5138	0.0871	87.86	95.19	97.71
	UniFuse [10]	0.2814	0.1063	0.4941	0.0701	88.97	96.23	98.31
	ACDNet [11]	0.2670	0.1010	0.4629	0.0646	90.00	96.78	98.76
	OmniFusion [13]	×	0.0900	0.4261	0.1483	91.89	**97.97**	**99.31**
	我们的方法（基础版）	0.2346	0.0906	0.4153	0.0577	91.92	97.55	99.12
	我们的方法（融合版）	**0.2263**	**0.0865**	**0.4052**	**0.0559**	**92.64**	97.68	99.11
	△AdaBins [4]	0.2560	0.0958	0.4484	0.1426	90.54	97.05	98.92
S2D3D	FCRN* [31]	0.3428	0.1837	0.5774	0.1100	72.30	92.07	97.31
	SliceNet* [6]	0.1757	0.0995	0.3509	0.0801	90.29	96.26	98.44
	我们的方法（基础版）*	0.1745	0.0919	**0.3229**	0.0593	92.12	**98.14**	**99.31**
	我们的方法（融合版）*	**0.1655**	**0.0843**	0.3232	**0.0575**	**92.90**	98.09	99.29
	Bifuse [9]	0.2343	0.1209	0.4142	0.0787	86.60	95.80	98.60
	HoHoNet [20]	0.2027	0.1014	0.3834	0.0668	90.54	96.93	98.86
	UniFuse [10]	0.2082	0.1114	0.3691	0.0721	88.97	96.23	98.31
	ACDNet [11]	0.1870	0.0984	0.3410	0.0664	88.72	97.04	98.95
	OmniFusion [13]	×	0.0950	0.3474	0.1599	89.88	97.69	**99.24**
	我们的方法（基础版）	**0.1760**	0.0936	**0.3272**	**0.0600**	91.53	**98.30**	99.18
	我们的方法（融合版）	0.1774	**0.0903**	0.3383	**0.0600**	**91.91**	97.82	99.12

表 VI：Pano3D [7] 数据集上的结果。“w”表示使用 Pano3D 提供的加权方法进行重新加权，使误差均匀分布。没有 Pano3D 官方结果的指标以”×”标记。

数据集	方法	误差指标 ⇓				准确率指标 (%) ⇑			
		$wRMSE$	$wRMSLE$	$wAbsRel$	$wSqRel$	$\delta_{1.1}^{ico^6}$	$\delta_{1.25}^{ico^6}$	$\delta_{1.25^2}^{ico^6}$	$\delta_{1.25^3}^{ico^6}$
在 *M3D* 上训练 在 GV2 tiny 上测试	UNet^{vnl}	0.5794	0.1247	0.2151	×	31.98	62.05	×	×
	ResNet$_{skip}^{comp}$	0.4993	0.1273	0.1758	×	40.78	80.31	×	×
	我们的方法（基础版）	**0.3654**	0.0800	0.1458	0.0675	45.44	80.93	96.03	99.13
	我们的方法（融合版）	0.3682	**0.0722**	**0.1353**	**0.0615**	**47.82**	**85.23**	**97.45**	**99.43**
在 *M3D* 上训练 在 GV2 medium 上测试	UNet^{vnl}	0.5901	0.1291	0.2269	×	31.21	61.02	×	×
	ResNet$_{skip}^{comp}$	0.4528	0.1618	0.1664	×	42.03	81.91	×	×
	我们的方法（基础版）	0.3502	0.0806	0.1500	0.0684	44.78	80.67	95.72	99.00
	我们的方法（融合版）	**0.3472**	**0.0729**	**0.1393**	**0.0614**	**46.95**	**84.66**	**97.16**	**99.33**
在 *M3D* 上训练 在 GV2 fullplus 上测试	UNet^{vnl}	0.8772	0.1769	0.2730	×	22.46	46.09	×	×
	ResNet$_{skip}^{comp}$	0.6607	0.2308	0.1836	×	**41.18**	**74.77**	×	×
	我们的方法（基础版）	0.5560	0.1255	0.2000	0.1430	33.27	62.20	84.66	94.05
	我们的方法（融合版）	**0.5243**	**0.1106**	**0.1804**	**0.1264**	37.35	67.99	**88.47**	**95.66**
在 *GV2 medium* 上训练 在 GV2 fullplus 上测试	OmniFusion [13]	0.6159	0.1054	0.1618	0.1418	42.87	73.39	91.36	96.40
	UniFuse [10]	0.4498	0.0894	0.1434	0.1110	48.11	79.47	93.89	97.74
	我们的方法（融合版）	**0.4479**	**0.0740**	**0.1318**	**0.0764**	**51.25**	**84.49**	**97.00**	**99.03**

采用 UniFuse [10] 中的技术。所有模型均在 8 块 V100 GPU 上训练，批量大小为 32。我们使用 AdamW [34]优化器，学习率为 $\gamma = 0.0003$。采用间隔为 3 个 epoch、衰减率为 0.9 的阶梯衰减学习率调度器，以加快训练过程。我们使用 5 个 epoch 的预热阶段，起始学习率为 0.00006，以获得更稳定的训练收敛效果。

C. 总体准确率

在真实世界数据集上，所提方法总体上表现最佳。如表 V 所示，所提方法在 S2D3D 上的 $AbsRel$ 降低了 **5.0%**（融合），在 M3D 上降低了**3.9%**（融合）；而在 $RMSE$ 分别为**4.2%**（基础）和**4.9%**（融合）。我们还将所提方法与平面图像方法 AdaBins [4]进行了比较；所提方法的结果更好，因为它在处理畸变和提取全局特征方面表现更佳。值得注意的是，与 ACDNet 相比，AdaBins 在 M3D 数据集上取得了更好的结果，这是因为 AdaBins 使用的自注意力模块大幅增强了其提取具有全局上下文的特征的能力。在 S2D3D 数据集上，融合 方法并非在所有评估指标上都优于基础 方法，因为 S2D3D 的训练图像数量较少（仅 1040 张），导致 CAF 模块训练不足。

在 Pano3D 上，根据表 VI 中的结果，与最佳基线结果相比，我们的方法将 $wAbsRel$ 降低了**23。0%**

表 VII：使用 M3D 数据集 [27]对不同骨干网络在最终准确率方面进行比较。我们使用 EffNet B5 表示 AdaBins [4]，中采用的 EffcientNet B5，并使用“D”表示解码器的维度。延迟在一块 NVIDIA TESLA P100 16GB 上进行测试。平面方法用△标记。

方法	骨干网络	参数量	延迟 y	绝对相对误差	RMSE	$\delta < 1.25$
UniFuse [10]	ResNet18	30.3M	25.5ms	0.1063	0.4941	88.97
Base	ResNet18	**19.9M**	**17.3毫秒**	0.1061	0.4761	88.67
融合	ResNet18	33.9M	25.2毫秒	**0.0995**	**0.4614**	**90.31**
Base	ResNet34	**29.9M**	**47.1ms**	0.1040	**0.4680**	89.33
融合	ResNet34	43.0M	81.3ms	**0.0988**	0.4720	**89.97**
ACDNet [11]	ResNet50	87.0M	72.4ms	0.1010	0.4629	90.00
基础 D=256	ResNet50	**32.9M**	**58.6ms**	0.1023	0.4638	89.43
基础 D=400	ResNet50	45.6M	92.9ms	0.0999	0.4607	89.85
融合 D=256	ResNet50	45.3M	99.6ms	0.0978	0.4552	90.44
融合 D=512	ResNet50	88.7M	144.3ms	**0.0944**	**0.4448**	**91.07**
△AdaBins [4]	EffNet B5	78.0M	199.9ms	0.0958	0.4484	90.54
Base	EffNet B5	**39.2M**	**80.7ms**	0.0947	0.4412	91.11
融合	EffNet B5	53.2M	111.2 毫秒	**0.0911**	**0.4280**	**91.90**

在小型划分上为（融合），在中型划分上为**16.3%**（融合）。$wRMSE$ 在这两个划分上分别降低**42.1%**（融合）和**43.5%**（融合）。我们还使用 GV2 的中型数据进行训练，并在 fullplus 上进行测试，以进一步参考该方法的性能。Pano3D 上的结果表明，所提方法具有良好的泛化能力。所提方法在 fullplus 测试划分上的准确率指标表现较弱；这可能是因为关注图像的全局上下文会使该方法容易受到训练数据与测试数据之间较大差异的影响。我们认为，增加训练数据可以解决这一问题。

如表 VIII 所示，在合成数据集上，所提方法整体上也表现最佳：与先前工作 UniFuse [10], 在 PSunCG 数据集上的结果相比，$RMSE$ 降低**6.5%**（融合），而 $AbsRel$ 降低**32.6%**（融合）。然而，在这两个数据集上，融合 并未显著优于基础 。这是因为三维模型渲染得到的深度与彩色图像的亮度之间存在局部相关性 [10], ，这促使网络将“局部”亮度信息作为预测深度的主要线索，因此融合 版本的全局上下文检索能力并未对最终结果有所贡献。

D. 消融研究

a) 不同骨干网络： 我们将所提方法应用于不同的骨干网络，并在表 VII 中展示准确率对比。将所提方法的骨干网络从 Swin 改为 ResNet 后，性能明显下降，但其表现仍优于 ACDNet 和 UniFuse 等方法，同时参数量更少、延迟更低。这证明，所提方法的提升不仅源于所使用的骨干网络，也得益于球面上的解码器以及我们的交叉注意力融合模块。

b) 室外数据集： 我们使用自行划分的训练集/测试集，评估了该方法在 KITTI 室外数据集上的性能 [35]，结果如表 IX 所示。尽管所提方法主要针对室内场景的深度估计，但其捕获全局上下文的能力有助于取得总体最佳结果。

c) 与重力方向未能完美对齐： 当相机未放置在水平面上时，可能会出现这种情况。表 X 中的结果展示了在全景图像轴与重力方向之间存在不同倾斜角时，所提方法的性能下降情况；评估所用的真实值是通过对 M3D 数据集的深度图像重新渲染获得的。对于倾斜角 2° 和 5°，所提方法的$RMSE$ 分别增加 **5。6%** 和 **18。3%**，而 SliceNet [6]的性能分别下降 **14。6%** 和 **25。9%**。这证明，所提方法具有更强的能力，能够抵御重力对齐不完美带来的影响。

V. 结论

本文提出了一种新颖的全景深度估计方法。通过在单位球面上采用均匀采样，并利用自注意力计算增强网络获取全局上下文的能力，该方法在所有公开数据集上都取得了出色的准确率。不过，所提方法仍有进一步改进的空间：将编码器部分移至单位球面上，可以增强方法的畸变去除能力，从而提高所提方法的准确率；在$\mathbf{f}_{\Pi\to\mathbf{S}^2}$ 和$\mathbf{f}_{\mathbf{S}^2\to\Pi}$ 的插值过程中，可以使用边缘保持滤波器，以进一步提升边缘处的深度估计质量。

参考文献

[1] P. Persson、L. ¨Ostr¨om、C. Olsson 和 K. ˚Astr¨om，“单目深度预测中的歧义参数化”，载于*3DV*，2021 年，第 761–770 页。[2] A. Mertan、D. J. Duff 和 G. Unal，“单幅图像深度估计：综述”，数字信号处理，第 103441 页，2022 年。[3] L. Huynh、P. Nguyen-Ha、J. Matas、E. Rahtu 和 J. Heikkila，“利用深度注意力体引导单目深度估计”，*arXiv* 预印本 *arXiv:2004.02760*，2020 年。[4] S. F. Bhat、I. Alhashim 和 P. Wonka，“Adabins：使用自适应分箱进行深度估计”，载于*CVPR*，2021 年，第 4009–4018 页。[5] R. Ranftl、A. Bochkovskiy 和 V. Koltun，“用于密集预测的视觉变换器”，载于*ICCV*，2021 年，第 12 179–12 188 页。[6] G. Pintore、M. Agus、E. Almansa、J. Schneider 和 E. Gobbetti，“SliceNet：使用基于切片的表示从单幅室内全景图进行深度密集估计”，载于*CVPR*，2021 年，第 11 536–11 545 页。[7] G. Albanis、N. Zioulis、P. Drakoulis、V. Gkitsas、V. Sterzentsenko、F. Alvarez、D. Zarpalas 和 P. Daras，“Pano3D：用于 360 度深度估计的整体基准与可靠基线”，载于*CVPR*，2021 年，第 3727–3737 页。[8] C. Sun、C.-W. Hsiao、N.-H. Wang、M. Sun 和 H.-T. Chen，“通过分而治之实现室内全景平面三维重建”，载于*CVPR*，2021 年，第 11 338–11 347 页。[9] F.-E. Wang、Y.-H. Yeh、M. Sun、W.-C. Chiu 和 Y.-H. Tsai，“Bifuse：通过双投影融合进行单目 360 度深度估计”，载于*CVPR*，2020 年，第 462–471 页。[10] H. Jiang、Z. Sheng、S. Zhu、Z. Dong 和 R. Huang，“UniFuse：用于 360 度全景深度估计的单向融合”，*IEEE*《机器人与自动化快报》，第 6 卷，第 2 期，第 1519–1526 页，2021 年。[11] C. Zhuang、Z. Lu、Y. Wang、J. Xiao 和 Y. Wang，“ACDNet：用于单目全景深度估计的自适应组合空洞卷积”，*arXiv* 预印本 *arXiv:2112.14440*，2021 年。[12] N. Zioulis、A. Karakottas、D. Zarpalas 和 P. Daras，“OmniDepth：室内球形全景的密集深度估计”，载于*ECCV*，2018 年，第 448–465 页。[13] Y. Li、Y. Guo、Z. Yan、X. Huang、Y. Duan 和 L. Ren，“OmniFusion：通过几何感知融合进行 360 度单目深度估计”，*arXiv* 预印本 *arXiv:2203.00838*，2022 年。[14] H. Du、H. Cao、S. Cai、J. Yan 和 S. Zhang，“球面变换器：将球面信号适配至卷积神经网络”，*arXiv* 预印本 *arXiv:2101 03848*，2021 年。

表 VIII：我们的方法与目前可用方法在合成数据集上的比较。我们使用 * 标记通过对齐方法评估的结果。使用 PSunCG 数据集训练的大多数方法只能以分辨率512 × 256估计深度图，而所提方法能够生成更高分辨率的深度图（1024 × 512），相关结果以 ◇标记。

数据集	方法	误差指标 ⇓ MAE	$AbsRel$	$RMSE$	$RMSE_{log}$	准确率指标 ⇑ $\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
PSunCG	OmniDepth* [12]	0.1624	0.1143	0.3710	0.0882	87.05	93.65	96.50
	FCRN* [31]	0.1346	0.0979	0.3973	0.0692	92.23	96.59	98.19
	我们的方法（基础版）*	0.0708	0.0355	**0.2639**	0.0346	97.64	98.94	99.46
	我们的方法（融合版）*	**0.0696**	**0.0340**	0.2646	**0.0341**	**97.81**	**98.98**	**99.48**
	Bifuse [9]	0.0789	0.0592	0.2596	0.0443	95.90	98.38	99.07
	UniFuse [10]	0.0765	0.0485	0.2802	0.0416	96.55	98.46	99.10
	我们的方法（基础版）	0.0647	0.0347	**0.2628**	0.0351	97.70	98.89	99.37
	我们的方法（融合版）	**0.0631**	**0.0327**	0.2631	**0.0346**	**97.82**	**98.92**	**99.38**
	◇我们的方法（基础版）	0.0572	0.0320	0.2220	0.0309	97.97	99.04	99.46
	◇我们的方法（融合版）	**0.0553**	**0.0298**	**0.2217**	**0.0303**	**98.09**	**99.07**	**99.47**
3D60	OmniDepth* [12]	0.1706	0.0931	0.3171	0.0725	90.92	97.02	98.51
	FCRN* [31]	0.1381	0.0699	0.2833	0.0473	95.32	99.05	99.66
	我们的方法（基础版）*	0.0766	0.0331	0.1802	0.0263	99.08	**99.71**	**99.87**
	我们的方法（融合版）*	**0.0733**	**0.0316**	**0.1770**	**0.0257**	**99.13**	**99.71**	**99.87**
	Bifuse [9]	0.1143	0.0615	0.2440	0.0428	96.99	99.27	99.69
	UniFuse [10]	0.0996	0.0466	0.1968	0.0315	98.35	99.65	99.87
	OmniFusion [13]	×	0.0430	**0.1808**	0.0735	98.59	99.69	**99.89**
	我们的方法（基础版）	0.0819	0.0367	0.1823	0.0271	99.05	**99.71**	99.87
	我们的方法（融合版）	**0.0809**	**0.0357**	0.1818	**0.0270**	**99.11**	**99.71**	99.87

表 IX：使用 KITTI 室外全景数据集进行的准确率测试结果 [35]。评估方法改编自 M3D 数据集 [27] 的评估方法。

方法	误差 ⇓ $AbsRel$	$RMSE$	$RMSLE$	准确率（%）⇑ $\delta < 1.25$
OmniFusion [13]	0.0977	4.536	0.193	90.80
UniFuse [10]	0.0945	4.186	0.076	92.58
所提方法	**0.0657**	**3.822**	**0.060**	**93.60**

表 X：全景图像未与重力方向最佳对齐时的误差角影响。不同倾斜角的数据通过对 M3D 数据集中的深度图像重新渲染获得 [27]。

方法	Tilt 角度	误差 ⇓ $AbsRel$	$RMSE$	$RMSLE$	准确率（%）⇑ $\delta < 1.25$
Bifuse [9]	0°	0.2048	0.6259	0.1134	84.52
	2°	0.3888	0.9805	0.1852	61.44
	5°	0.4905	1.0225	0.2250	54.40
SliceNet [6]	0°	0.1764	0.6133	0.1045	87.16
	2°	0.2645	0.7026	0.1334	72.56
	5°	0.3032	0.7720	0.1482	68.79
Ours	0°	0.0865	0.4052	0.0559	92.64
	2°	0.0927	0.4282	0.0600	91.60
	5°	0.1091	0.4792	0.0711	88.14

[15] T. S. Cohen、M. Geiger、J. K¨ohler 和 M. Welling，“球面卷积神经网络”，载于国际学习表征会议，2018 年。[16] K. M. Gorski、E. Hivon、A. J. Banday、B. D. Wandelt、F. K. Hansen、M. Reinecke 和 M. Bartelmann，“Healpix：一种用于球面分布数据的高分辨率离散化与快速分析的框架”，天体物理学杂志，第 622 卷，第 2 期，第 759 页，2005 年。[17] Y. Wang、X. Li、M. Shi、K. Xian 和 Z. Cao，“面向移动设备的快速、准确单目深度估计的知识蒸馏”，载于*CVPRW*，2021 年，第 2457–2465 页。[18] Y. Chen、H. Zhao、Z. Hu 和 J. Peng，“用于单目深度估计的基于注意力的上下文聚合网络”，《国际机器学习与控制论杂志》，第 12 卷，1596，2021 年 6，1583 期，第 – 页。[19] J. Watson、O. Mac Aodha、V. Prisacariu、G. Brostow 和 M. Fir-man，“The temporal opportunist：自监督多帧单目深度”，*arXiv* 预印本 *arXiv:2104.14540*，2021 年。[20] C. Sun、M. Sun 和 H.-T. Chen，“HoHoNet：利用潜在水平特征实现 360 度室内整体理解”，载于*CVPR*，2021 年，第 2573–2582 页。[21] Y. Su 和 K. Grauman，“用于紧凑型球面卷积的核变换器网络”，载于*CVPR*，2019 年，第 9442–9451 页。[22] C. Fernandez-Labrador、J. M. F´acil、A. Perez-Yus、C. Demonceaux，

J. Civera 和 J. J. Guerrero，“用于布局的角点：从360度图像端到端恢复布局”，*arXiv:1903.08094*，2019。[23] N. Wang、B. Solarte、Y. Tsai、W. Chiu 和 M. Sun，“360sd-net：利用可学习代价体进行360°立体深度估计”，载于*ICRA*，2020年，第582–588页。[24] R. Liu、J. Lehman、P. Molino、F. P. Such、E. Frank、A. Sergeev 和 J. Yosinski，“卷积神经网络一个耐人寻味的缺陷及其coordconv解决方案”，载于*NeurIPS*，2018年，第9628–9639页。[25] S. Cho、R. Jung 和 J. Kwon，“球面变换器”，*arXiv* 预印本 *arXiv:2202.04942*，2022年。[26] Y. Lee、J. Jeong、J. Yun、W. Cho 和 K.-J. Yoon，“Spherephd：利用非欧几里得球面多面体表示在 360° 图像上应用卷积神经网络”，《*IEEE*模式分析与机器智能汇刊》，第44卷，第2期，第834–847页，2022年。[27] A. Chang、A. Dai、T. Funkhouser、M. Halber、M. Niessner、M. Savva、S. Song、A. Zeng 和 Y. Zhang，“Matterport3d：从室内环境中的RGB-D数据中学习”，*arXiv*预印本 *arXiv:1709.06158*，2017年。[28] X. Chu、Z. Tian、Y. Wang、B. Zhang、H. Ren、X. Wei、H. Xia 和 C. Shen，“Twins：重新审视视觉变换器中的空间注意力设计”，*NeurIPS*，第34卷，2021年。[29] A. Alvarez-Gila、J. van de Weijer 和 E. Garrote，“利用对抗网络从RGB图像重建空间上下文感知光谱图像”，载于*ICCV*，2017年，第480–490页。[30] I. Armeni、S. Sax、A. R. Zamir 和 S. Savarese，“用于室内场景理解的联合2D-3D语义数据”，*arXiv*预印本 *arXiv:1702 01105*，2017.。[31] I. Laina、C. Rupprecht、V. Belagiannis、F. Tombari 和 N. Navab，“利用全卷积残差网络进行更深层次的深度预测”，载于*3DV*，2016年，第239–248页。[32] F.-E. Wang、H.-N. Hu、H.-T. Cheng、J.-T. Lin、S.-T. Yang、M.-L. Shih、H.-K. Chu 和 M. Sun，“从 360° 视频中自监督学习深度和相机运动”，载于*ACCV*，2018年，第53–68页。[33] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo，“Swin变换器：采用移位窗口的分层视觉变换器”，载于*ICCV*，2021年，第10 012–10 022页。[34] I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，载于第七届国际学习表征会议（*ICLR 2019*），美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。[35] G. P. de La Garanderie、A. A. Abarghouei 和 T. P. Breckon，“消除盲点：将3D目标检测和单目深度估计适配于360度全景图像”，载于*ECCV*，第11217卷。施普林格，2018年，第812–830页。