DOI: 10.19650/j.cnki.cjsi.J2413429

基于自监督学习的热成像与激光雷达融合深度补全方法*

于 睿[1],马国梁[2],郭 健[1],许立松[1]
(1.南京理工大学自动化学院 南京 210094; 2.南京理工大学能源与动力工程学院 南京 210094)

摘 要:深度补全是一种利用稀疏深度数据生成高分辨率稠密深度图的环境感知技术。然而,现有深度补全算法在昏暗或低照度场景中预测深度图的准确度不足,在极端光照条件下的应用效果较差。针对该问题,提出一种基于自监督深度学习的热成像与激光雷达融合深度补全方法,用于训练网络模型在低光照或无光照的条件下生成像素级稠密的深度图。所提网络为编码器-解码器架构,以热图像和激光雷达的稀疏深度图作为编码器输入,在不同图像尺度上进行特征融合,解码器逐层对融合后的特征进行上采样和深度预测,生成稠密深度图。其次,设计了基于自注意力与跨注意力机制的多模态融合模块嵌入到编码器,通过自适应加权增强特征融合效果,提升预测稠密深度图的准确度。最后,构建了自监督学习框架,利用温度重建损失和稀疏深度损失进行自监督训练,无需额外的深度真值标注过程。在公开数据集上的实验验证表明,所提方法在不同光照条件下均能稳定生成稠密深度图。相较于现有深度补全基准方法,平均绝对误差在 MS2 和 VIVID 数据集上分别降低了 44.49% 和 25.28%。在低光或无光环境下,通过融合热成像与激光雷达数据的互补优势,显著提高了深度预测的准确性和稳健性,为低光照场景下的环境感知提供了有效解决方案。

关键词:深度图补全;多传感器数据融合;热成像;自监督学习;环境感知
中图分类号:TH811 TP242 **文献标识码**:A **国家标准学科分类代码**:510.4050

Self-supervised learning-based depth completion method using thermal imaging and LiDAR fusion

Yu Rui[1], Ma Guoliang[2], Guo Jian[1], Xu Lisong[1]
(*1. School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China*;
2. School of Energy and Power Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Depth completion is a technique for generating high-resolution dense depth maps from sparse depth data for environmental perception. Existing methods struggle with accuracy in low-light or dark conditions, performing poorly under extreme lighting. This article proposes a self-supervised method that fuses thermal images and LiDAR data to complete dense depth maps in low-light or no-light scenarios. The network adopts an encoder-decoder structure, using thermal images and sparse LiDAR depth as inputs. Features are fused at multiple scales in the encoder, and the decoder upsamples them to predict dense depth maps. Multi-modal fusion modules based on self-attention and cross-attention are embedded in the encoder to enhance feature fusion with adaptive weighting, improving prediction accuracy. A self-supervised framework is established with temperature reconstruction and sparse depth losses, removing the need for depth ground truth. Experiments on public datasets show that the method generates dense depth maps stably under various lighting conditions. Mean absolute error decreases by 44.49% on MS2 and 25.28% on VIVID compared to benchmarks. By leveraging thermal and LiDAR data's complementary strengths, this method improves depth prediction accuracy and robustness in low-light environments, offering an effective solution for perception in challenging lighting.

Keywords: depth completion; multi-sensor data fusion; thermal imaging; self-supervised learning; environmental perception

收稿日期:2024-10-31 Received Date: 2024-10-31
*基金项目:江苏省农业科技自主创新资金项目(CX(24)1023)、中央高校基本科研业务费专项(2024301002)资助

0　引　　言

环境感知是机器人巡检[1]和自动驾驶领域的研究热点,其中深度估计与深度补全[2-3]作为核心任务,在三维场景理解任务中发挥重要作用。深度估计旨在通过相机推断像素距离信息,重建场景三维结构。深度补全则侧重对激光雷达点云等稀疏深度数据进行稠密化处理,生成完整的像素级深度图,填补数据空洞。深度补全作为自主系统精确的环境感知技术,为移动机器人位姿估计、导航和路径规划决策[4-6]等下游任务提供了基础支持。

基于机器学习的深度补全算法通过神经网络实现端到端场景几何推断,受到广泛关注。常用的可见光(red green blue, RGB)相机在光照良好时能捕捉丰富的纹理信息,但在夜间等低光照条件下易受噪声和运动模糊影响,导致深度估计的准确度下降。激光雷达作为主动式传感器则不受光照条件限制,能稳定测量距离信息,但点云稀疏性难以展现场景的结构和细节。

为克服单一传感器的局限性,现有研究尝试融合 RGB 图像与激光雷达数据[7],以提升场景感知的完整性。多传感器融合的深度补全方法可分为有监督和自监督两类。有监督方法通过最小化相对于深度真值的损失函数,将稀疏深度图和图像映射为稠密深度图。早期方法基于压缩传感或形态学算子[8],近期研究则优化了网络架构,如联合卷积注意力的金字塔结构[9]。Yu 等[10]设计了点云处理网络聚合 2D 与 3D 点云特征重建缺失的深度信息。Tang 等[11]使用了一种三阶段融合的策略,避免了直接对稀疏的深度图卷积。Shi 等[12]采取了 RGB 图像与深度图直接拼接后提升至三维空间再融合的策略。Yan 等[13]则分解三维稀疏点云到不同二维空间由三视图融合深度补全。

尽管有监督方法在数据集上表现良好,但受限于稠密深度图的获取难度,其实际部署仍存在局限性。为此,自监督深度补全方法通过降低对深度真值的依赖,利用双目相机或单目视频生成额外信息,通过最小化光度差值完成训练[14-15]。一些方法引入额外的稀疏深度图以构建时空一致性[16]来优化深度预测结果。Jeon 等[17]利用视觉同步定位与建图技术提取线条特征改善估计结果。Yan 等[18]设计了尺度估计网络将深度图预测与尺度估计分离以获得更好的视觉效果。部分研究者则利用了相邻帧间的几何信息[19]提高了深度补全整体性能,或设计了校准反投影层[20]优化模型框架,进一步提升深度预测网络性能。

虽然 RGB 激光雷达融合的深度补全方法在白天光照良好时能生成结构清晰、测距准确的稠密深度图,但在低光或无光环境下表现较差,因 RGB 相机在弱光条件下的局限性直接影响了补全效果。相比之下,热成像相机作为一种不受光照变化影响的被动传感器,可在无光环境中测量物体表面温度分布,具备独特的三维环境感知优势[21]。然而热成像在深度补全领域的研究仍显不足,其与激光雷达数据的融合应用未被充分挖掘,限制了其在极端光照条件下的感知潜力。

综上分析,提出一种基于热成像与激光雷达融合的自监督深度补全方法。该方法通过多帧热图像和稀疏深度图训练深度图预测网络,在测试部署时,仅需单帧热图像和稀疏深度图即可生成稠密深度图。

本研究方法贡献如下:1) 提出了一种深度补全的自监督深度学习方法,利用温度重建损失和稀疏深度损失驱动网络模型训练,无需深度真值。测试部署时网络可以从热图像和稀疏深度图中计算出与热图像对应的稠密深度图。2) 为有效融合激光雷达数据和热图像数据,设计了 2 个融合模块,分别用于融合多模态数据和深度图-热图像数据。这些模块嵌入在网络的编码器中,符合整体网络模型的层内融合策略。3) 通过热图像重映射方法,增加了热图像的纹理细节和结构信息,以增强相邻帧之间的视觉差异性,从而优化网络训练结果,提升稠密深度图的预测准确性。4) 在 2 个公开的室外数据集上进行了广泛的验证实验,结果表明使用提出的网络模型能够在低光照或几乎无光照的条件下,可靠且稳健地预测稠密深度图,相较于传统的 RGB 方法性能显著提升。

1　系统介绍

1.1　系统框架

本研究方法的系统框架如图 1 所示。深度图预测网络接收热图像和稀疏深度图作为输入,编码器提取并逐级融合两支数据特征,解码器生成稠密深度图。位姿估计网络通过相邻帧热图像预测相机运动位姿,辅助完成自监督训练过程。

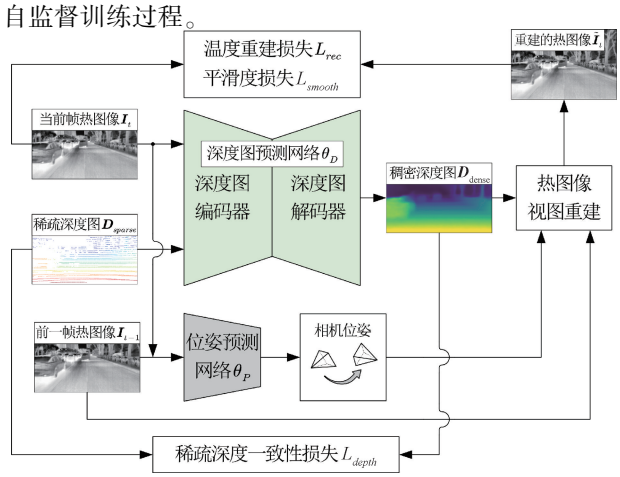


图 1　所提方法系统框架

Fig. 1　The proposed system framework

提出的方法基于热图像视图重建机制进行自监督训练,利用热图像的投影关系重建当前帧热图像,并通过自监督损失驱动网络优化。本研究方法引入了稀疏深度图与预测深度图的深度一致性损失函数,用于保持预测深度距离尺度一致性。

此外,通过热图像重映射和图像增强技术,优化热图像的细节和结构信息,改善其纹理较弱的特性,用于增强视觉差异性提升训练效果。

1.2 网络模型

图 2 展示了网络模型的整体结构,分别为深度图预测网络(由深度图编码器和深度图解码器构成)和仅作为辅助训练的位姿预测网络。

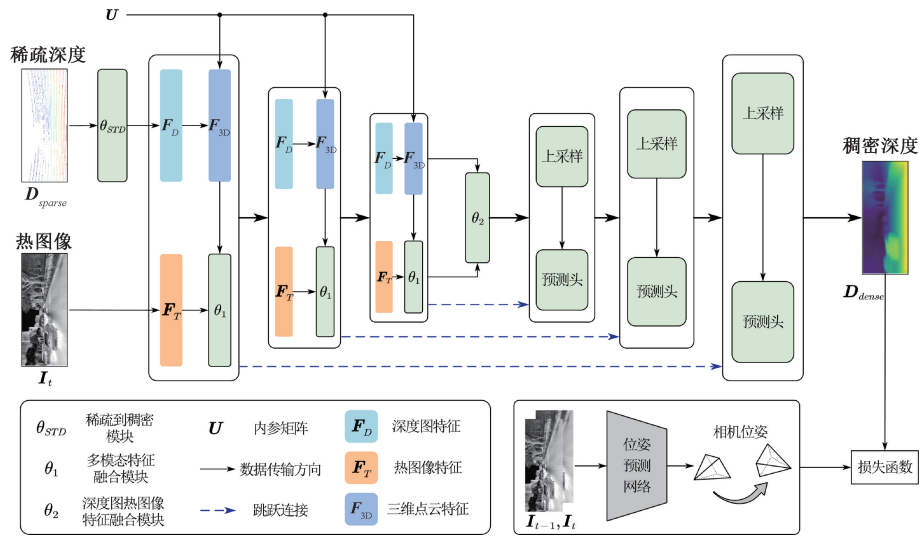


图 2 所提网络模型整体结构

Fig. 2 Overall architecture of the proposed network model

1)深度图编码器

深度图编码器从输入的稀疏深度图和热图像中提取并融合特征,生成多尺度特征图以支持解码器输出稠密深度图。由于点云数量远少于热图像像素,稀疏深度图中存在大量 0 值区域,限制了编码器的特征提取能力,影响层内数据融合效果。

为此,采用轻量级的稀疏到稠密(sparse to dense, STD)模块对稀疏深度图进行预稠密化。STD 模块由不同核尺寸的最大池化层、最小池化层和卷积层网络组成。最大池化保留远处结构但易丢失近处细节,最小池化则保留近距离细节并抑制远距离特征。该对偶设计在预稠密化过程中尽可能保留图像细节。随后使用 4 个可学习卷积层整合最大池化与最小池化的结果,生成半稠密深度图。

编码器为多尺度结构,在单个编码器层内实现多模态特征的提取与融合。如图 3 所示,单层编码器接收热图像特征图、深度特征图、相机内参矩阵及融合特征图为输入,分别通过独立分支处理后传递至下一层,最终在最后一层通过深度图-热图像融合模块(depth-thermal image fusion, DTF)完成特征融合。

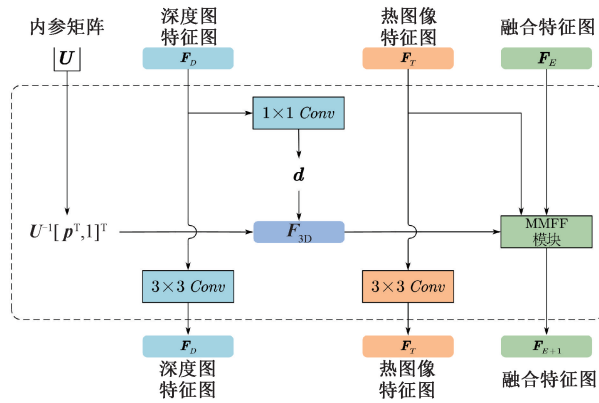


图 3 单个深度图编码器层的网络结构

Fig. 3 Network architecture of a single depth encoder layer

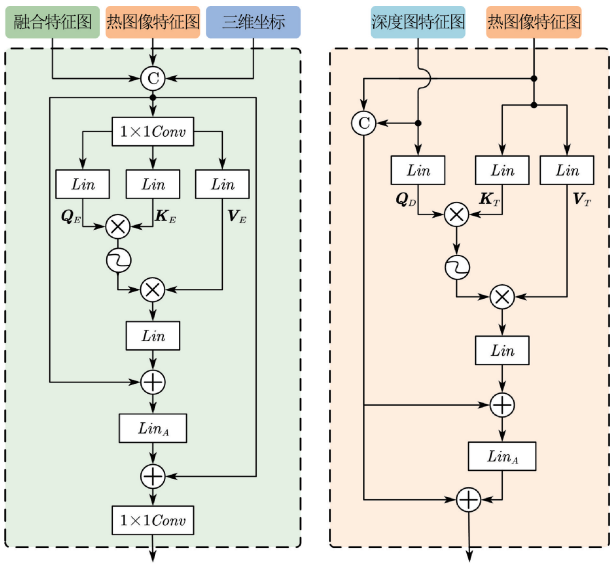
对于每个编码器层内的数据融合,首先将热图像特征图的像素坐标转为齐次坐标,并通过热成像相机的内参矩阵逆投影像素到归一化图像平面,从而将热图像的所有像素提升至三维空间。随后,利用一个 1×1 卷积层将深度特征图压缩为单通道的深度图 $\boldsymbol{d}$。利用 $\boldsymbol{d}$ 恢复三维空间下热图像像素的尺度信息,将像素坐标转移至三维相机坐标系,生成热图像的三维坐标特征图 $\boldsymbol{F}_{3D}$。

热图像特征图、三维坐标特征图及来自前一层融合特征图在通道维度拼接,形成多模态特征张量。设计了一种基于自注意力机制的多模态特征融合模块(multi-

modality feature fusion, MMFF)用于处理该张量,生成的融合特征图传递至下一编码器层,并通过跳跃连接输出给解码器,避免梯度消失和性能退化。

2)多模态数据融合

多模态数据特征尺度不一致,直接融合容易导致模态间的特征权重分布失衡与信息丢失。为此,提出了MMFF 和 DTF 融合模块,如图 4 所示,分别基于自注意力机制和交叉注意力机制实现特征自适应调整融合。



(a) 多模态特征融合模块
(a) Multi-modality feature fusion module

(b) 深度图-热图像融合模块
(b) Depth-thermal image fusion module

图 4　多模态特征融合模块和深度图-热图像融合模块的网络结构

Fig. 4　Network architecture of multi-modality feature fusion module and depth-thermal image fusion module

MMFF 模块用于不同模态数据在编码器内部的层间融合,通过学习不同模态特征之间的关系与相似度,动态调整每个融合权重,生成综合特征表达。如图 4(a)所示,该模块接收形状为 (h,w,c) 的多模态特征堆叠张量 $\boldsymbol{F}_E$,E 表示当前的编码器层序号,h,w 为当前层对应的特征图尺寸,c 为该层的通道数量。经过一个 1 × 1 卷积层处理后,特征张量 $\boldsymbol{F}_E$ 的形状从 (h,w,c) 变换为 (hw,c)。为了捕捉模态内部特征点之间的长距离依赖关系,通过3 个不同的权重张量将 $\boldsymbol{F}_E$ 线性变换为 3 个不同的张量。

$$\boldsymbol{Q}_E = \boldsymbol{F}_E\boldsymbol{W}_E^Q, \boldsymbol{K}_E = \boldsymbol{F}_E\boldsymbol{W}_E^K, \boldsymbol{V}_E = \boldsymbol{F}_E\boldsymbol{W}_E^V \tag{1}$$

其中,$\boldsymbol{Q}_E$,$\boldsymbol{K}_E$,$\boldsymbol{V}_E$ 分别代表 MMFF 模块下的查询张量、键张量和值张量 $\boldsymbol{W}_E^Q$,$\boldsymbol{W}_E^K$,$\boldsymbol{W}_E^V$ 为其对应的权重张量。该线性变换操作不改变张量 $\boldsymbol{F}_E$ 的形状。对 $\boldsymbol{Q}_E$ 和 $\boldsymbol{K}_E$ 计算点积注意力,得到自注意力图。

$$\boldsymbol{M}_E = Softmax(\boldsymbol{Q}_E\boldsymbol{K}_E^{\mathrm{T}}) \tag{2}$$

通过 $Softmax(\cdot)$ 函数对自注意力图进行归一化,确保每个元素的注意力权重总和为 1。自注意力图反映了每个元素与其他元素的相似程度,引导模型在特征融合时对不同元素的权重进行调整。将 $\boldsymbol{V}_E$ 与自注意力图 $\boldsymbol{M}_E$ 相乘,得到自注意力输出量。

$$\widetilde{\boldsymbol{F}}_E = Attention(\boldsymbol{Q}_E,\boldsymbol{K}_E,\boldsymbol{V}_E) + \boldsymbol{F}_E \tag{3}$$

其中,$Attention(\boldsymbol{Q}_E,\boldsymbol{K}_E,\boldsymbol{V}_E) = \boldsymbol{M}_E\boldsymbol{V}_E$。增强后的张量 $\widetilde{\boldsymbol{F}}_E$ 经过层归一化、线性变换、激活函数处理,并通过一个 1 × 1 卷积层网络生成后一层级的融合特征张量。

$$\boldsymbol{F}_{E+1} = Conv(Lin_A(Lin(LN(\widetilde{\boldsymbol{F}}_E)) + \boldsymbol{F}_E) \tag{4}$$

其中,$LN(\cdot)$、$Lin(\cdot)$ 和 $Conv(\cdot)$ 分别代表层归一化函数、线性层和卷积层,$Lin_A(\cdot)$ 表示带有激活函数的线性层。计算得到的新的融合特征图传输给后一个编码器层作为输入。

此外,在每个编码器层中 $\boldsymbol{Q}_E$ 与 $\boldsymbol{K}_E$ 之间的自注意力图计算过程导致了较高的计算复杂度,为加快模型的训练和推理速度,设计了一个简化的多模态特征融合模块。原始的线性变换将张量尺寸映射为 (hw,c),使用带状平均池化操作,将 $\boldsymbol{Q}_E$ 和 $\boldsymbol{K}_E$ 的尺寸分别压缩为 (w,c) 和 (c,h),自注意力图的计算复杂度从 $(hw)^2$ 降低至 hw。在实验分析中展示该简化模块对加速模型推理的优化效果。

如图 4(b)所示,经多个编码器层的提取特征,热图像数据分支和深度图数据分支的特征图在最后的编码器层使用基于交叉注意力机制的 DTF 模块进行全局特征融合。具体来说,DTF 模块接收来自两个分支的热图像特征图 $\boldsymbol{F}_T$ 与深度图特征图 $\boldsymbol{F}_D$,通过线性变换与点积计算获得交叉注意力输出:

$$\widetilde{\boldsymbol{F}}_{DT} = Attention(\boldsymbol{Q}_D,\boldsymbol{K}_T,\boldsymbol{V}_T) + \boldsymbol{F}_{DT} = \boldsymbol{M}_{DT}\boldsymbol{V}_T + \boldsymbol{F}_{DT} \tag{5}$$

其中,与式(1)类似,$\boldsymbol{Q}_D$ 来自于深度图特征图 $\boldsymbol{F}_D$ 的线性变换,$\boldsymbol{K}_T$,$\boldsymbol{V}_T$ 则来自于热图像特征,而 $\boldsymbol{M}_{DT}$ 则是交叉注意力图 $\boldsymbol{M}_{DT} = Softmax(\boldsymbol{Q}_D\boldsymbol{K}_T^{\mathrm{T}})$。$\boldsymbol{F}_{DT}$ 表示 $\boldsymbol{F}_D$ 与 $\boldsymbol{F}_T$ 的在通道维度的拼接。计算得到的交叉注意力输出量在增强其非线性性能后,与输入量进行残差连接:

$$\hat{\boldsymbol{F}}_{DT} = Lin_A(Lin(LN(\widetilde{\boldsymbol{F}}_{DT})) + \boldsymbol{F}_{DT} \tag{6}$$

最终深度图与热图像的融合特征 $\hat{\boldsymbol{F}}_{DT}$ 传输至解码器用于生成稠密深度图。

1.3　热图像重映射

热成像相机通常输出归一化的热图像或原始热图像。为适应人眼观察,默认设置的归一化热图像会根据当前帧温度测量进行最大值和最小值的重缩放与归一化处理。然而,此类重缩放会导致每帧图像的温度分布不一致,破坏了图像序列的时间一致性,不利于自监督训练。原始热图像的测温范围固定,保证了时间一致性,但

其常见温度分布较窄(如 10℃~45℃),与宽泛的测温范围(如-25℃~150℃)不匹配,导致原始图像的细节和对比度不足,图像差异度较小,难以有效驱动网络自监督训练。

为解决此问题,对原始热图像进行重映射,在保证时间一致性的同时增强图像细节,生成足够的图像差异性以驱动网络训练。首先,将图像序列的所有热图像限制在有效测温范围 $\{\tau_{\min},\tau_{\max}\}$,并对温度分布进行重映射,增强热图像信息表达。具体来说,统计图像序列中的当前帧与相邻帧构建温度值直方图。

$$hist(i)=n_i,\ i=1,2,\cdots,N_\tau \tag{7}$$

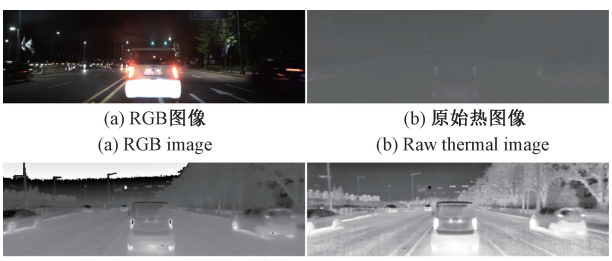
其中,N_τ 表示直方图根据测温范围划分的子区间数量,i 为区间序号,n_i 为子区间内的像素数量。对于原始热图像中第 i 个子区间内的温度测量值 τ,经过重映射后变换为:

$$\tilde{\tau}=\alpha_i\cdot\left(\frac{\tau-b_i}{b_{i+1}-b_i}\right)+\tilde{b}_i \tag{8}$$

其中,$\{b_i,b_{i+1}\}$ 是第 i 个子区间的温度最大值与最小值,α_i 代表了当前子区间内像素数量在整体像素数量中的占比,$\tilde{b}_i$ 是重映射后子区间新的偏移量,即:

$$\tilde{b}_i=\sum_{k=0}^{i-1}\alpha_k \tag{9}$$

热图像重映射操作压缩了直方图中像素数量较少的子区间,将有效测温区间扩展至全局范围,并对原始热图像进行归一化。全局温度裁剪与重映射基于整个训练图像序列进行操作,在训练中保持相邻帧之间测温的时间一致性。此外,使用限制对比度自适应直方图均衡化算法对热图像进一步优化细节,增强模糊热图像的对比度,处理效果如图 5 所示。



(a) RGB图像 (a) RGB image

(b) 原始热图像 (b) Raw thermal image

(c) 限制测温范围后的热图像 (c) Clipped thermal image

(d) 重映射后的热图像 (d) Remapped thermal image

图 5 热图像重映射效果

Fig. 5 Effect of thermal image remap ping

1.4 自监督学习和损失函数

不依赖于深度真值,本方法利用自动生成的监督信息计算损失驱动模型训练。如图 1 所示,自监督学习框架基于图像重建机制。给定图像序列中的相邻两帧热图像 $\{\boldsymbol{I}_{t-1},\boldsymbol{I}_t\}$ 和一张稀疏深度图 $\boldsymbol{D}_{spase}$,首先使用位姿估计网络生成两帧热图像之间的位姿 $\boldsymbol{P}_{t\to t-1}=\theta_P(\boldsymbol{I}_{t-1},\boldsymbol{I}_t)$,并使用深度图预测网络生成稠密深度图 $\boldsymbol{D}_{dense}=\theta_D(\boldsymbol{D}_{sparse},\boldsymbol{I}_t)$。通过图像重建操作,当前热图像帧可以由前一帧 $\boldsymbol{I}_{t-1}$ 重建生成。该图像重建过程基于视图投影[22]。

$$\tilde{\boldsymbol{I}}_t=\boldsymbol{I}_{t-1}\langle Proj(\boldsymbol{D}_{dense},\boldsymbol{P}_{t\to t-1},\boldsymbol{U})\rangle \tag{10}$$

其中,$\tilde{\boldsymbol{I}}_t$ 表示被重建的当前帧热图像,$Proj(\cdot)$ 为投影函数,将 $\boldsymbol{I}_t$ 的像素坐标通过稠密深度 $\boldsymbol{D}_{dense}$ 和热成像相机内参矩阵 $\boldsymbol{U}$ 投影到 $\boldsymbol{I}_{t-1}$。$\langle\cdot\rangle$ 表示采样算子,在前一帧热图像 $\boldsymbol{I}_{t-1}$ 上进行插值采样。利用重建热图像 $\tilde{\boldsymbol{I}}_t$ 与当前帧热图像 $\boldsymbol{I}_t$ 计算损失函数。

综上,给定热图像 $\boldsymbol{I}_t$ 与重建的热图像 $\tilde{\boldsymbol{I}}_t$,计算温度重建损失,其由结构相似性指数(structural similarity index measure, SSIM)损失[23]和 L1 损失构成。

$$L_{rec}=\frac{1}{2}\lambda_{SSIM}(1-SSIM(\boldsymbol{I}_t,\tilde{\boldsymbol{I}}_t))+\lambda_{tem}\|\boldsymbol{I}_t-\tilde{\boldsymbol{I}}_t\|_1 \tag{11}$$

其中,λ_{SSIM}、λ_{tem} 分别表示 SSIM 损失和 L1 损失的权重。

对于预测得到的稠密深度图 $\boldsymbol{D}_{dense}$ 和热图像 $\boldsymbol{I}_t$,为平滑 $\boldsymbol{D}_{dense}$ 中的不连续区域,采用边缘感知平滑度损失[24]。

$$L_{smooth}=\|\partial_x\boldsymbol{D}_{dense}\|_1 e^{-\|\partial_x\boldsymbol{I}_t\|_1}+\|\partial_y\boldsymbol{D}_{dense}\|_1 e^{-\|\partial_y\boldsymbol{I}_t\|_1} \tag{12}$$

其中,∂_x、∂_y 分别表示求解图像在 x 和 y 方向上梯度。

仅使用温度重建损失和平滑度损失将导致模型难以学习到真实场景的尺度信息,为解决该问题将输入的稀疏深度图 $\boldsymbol{D}_{sparse}$ 与预测得到的稠密深度图 $\boldsymbol{D}_{dense}$ 之间的差值作为监督信息,保持 $\boldsymbol{D}_{dense}$ 的场景尺度与 $\boldsymbol{D}_{sparse}$ 一致。该差值通过最小化 $L1$ 损失表示为:

$$L_{depth}=\|\boldsymbol{M}_{|\boldsymbol{D}>0|}\cdot(\boldsymbol{D}_{sparse}-\boldsymbol{D}_{dense})\|_1 \tag{13}$$

其中,$\boldsymbol{M}_{|\boldsymbol{D}>0|}$ 为稀疏深度图的掩膜,设置 $\boldsymbol{D}_{sparse}$ 深度值 >0 的像素为有效区域。

综上,本方法的总体损失函数为各项损失量的加和为:

$$L_{total}=L_{rec}+\lambda_S L_{smooth}+\lambda_D L_{depth} \tag{14}$$

其中,λ_S,λ_D 分别表示平滑度损失和稀疏深度一致性损失对应的权重。

2 实验及结果

2.1 数据集

如 Geiger 等[25]制作的经典深度补全数据集缺乏夜间场景和热成像数据,无法验证本方法的有效性。因此,

选用多光谱立体视觉(multi-spectral stereo, MS2)数据集[26]和可见性视觉(vision for visibility, VIVID)数据集[27]进行模型训练与评估。MS2 数据集是一个多模态公开数据集,涵盖激光雷达、惯性组合导航数据、RGB 图像、热图像等,采集于城市建筑和道路等典型场景。为评估所提网络在不同光照条件下的性能,从 MS2 数据集中选取 35 902 组样本作为训练集,8 184 组样本作为验证集,包含白天和夜间的图像数据。VIVID 数据集同样由多传感器采集,样本来源于室外手持式设备。本文选取其 3 243 组样本作为训练集,989 组样本作为测试集,涵盖白天与夜晚的室外场景。

2.2　实验设置细节

根据热成像相机和激光雷达参数,使用单通道热图像以及稀疏深度图进行实验,深度值测试范围设置为 0.1~100 m。实际场景实验设备如图 6 所示,使用手持式安装的激光雷达 Ouster OS1-32、热成像相机 FLIR A35 和迷你上位机 Intel NUC 11,搭载移动电源基于机器人操作系统(ROS)进行实机实验,算法部署在 Ubuntu 20.04 系统中,使用 Matplotlib 进行可视化结果显示。训练阶段,使用 ResNet18 网络模型[28]作为位姿估计器计算热图像序列 $\{\boldsymbol{I}_{t-1},\boldsymbol{I}_t,\boldsymbol{I}_{t+1}\}$ 间的相机位姿,结合稀疏深度图 $\boldsymbol{D}_{sparse}$ 和热成像相机内参矩阵,构建自监督学习框架驱动训练。测试阶段仅使用当前帧数据即可预测稠密深度图。设置权重 $\lambda_{L1}=0.15$,$\lambda_{SSIM}=0.95$,$\lambda_S=0.01$,$\lambda_D=0.8$,热图像重映射算法参数 $\{\tau_{\min},\tau_{\max}\}$ 设置为 $\{30.0,45.0\}$。模型在 Nvidia RTX 3080Ti GPU 上进行训练测试,使用 Adam 优化器,初始学习率为 0.000 1,批量大小为 8,在 MS2 数据集上训练 40 轮。

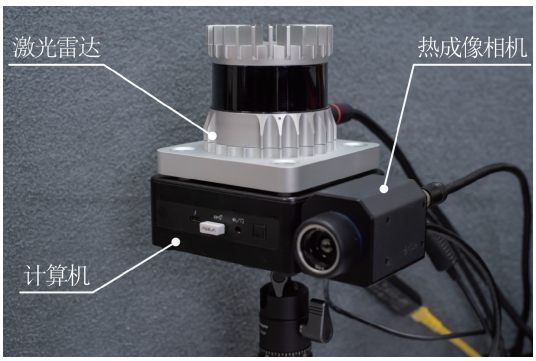


图 6　实际场景实验设备

Fig. 6　Experimental setup in real scenarios

2.3　评估标准

根据广泛应用于深度估计和深度补全领域的测试标准[19,29]对提出的标准进行实验结果的量化评估,度量标准包括平均绝绝对误差(mean absolute error, MAE)、均方根误差(root mean square error, RMSE)、绝对值相对误差(absolute relative error, Abs Rel)、平方相对误差(squared relative error, Sq Rel)、准确度指标等。模型预测每个测试样本的稠密深度图,并根据深度图和数据集提供的深度真值计算误差指标(越小越好)以及准确度指标(越大越好)。

2.4　MS2 数据集上的实验测试

为验证提出方法的有效性,将本方法与其他自监督深度补全方法进行对比。定量的实验测试结果如表 1 所示,已开源的对比实验方法 FusionNet[14]、VOICED[15]、KBNet[20]使用 RGB 图像和激光雷达作为输入源,而 FusionNet 中所提的另一算法 ScaffNet[14]仅以激光雷达作为输入源。由于 MS2 数据集缺乏完全稠密的深度图作为监督,因此本次实验中 ScaffNet 通过虚拟合成数据集[30]的预训练模型进行微调训练。对不同场景下的样本测试结果取平均后,提出的方法在各个误差指标和准确度指标取得了最优结果,而所提方法的快速版本,也即表 1 中的本文(快速),获得了次优的结果。图 7 展示了提出的方法与基准方法定性对比结果,而图 8 和 9 则是白天与夜晚情况下,预测的稠密深度图对比深度真值的误差图。误差图根据绝对值相对误差计算,图中误差较小的区域最小值为 0,误差较大区域最大值为 0.04。

在夜间由于可见光图像因曝光不足而变得较暗,同时伴随细节丢失,使得深度估计模型在多个区域产生错误,容易出现模糊和深度不连续现象。在图 7 的矩形框标识的区域,物体轮廓模糊且深度值不准确,从而导致出现伪影。而图 8 和 9 显示在全局和矩形框的局部区域所提方法均展现出一致的低误差性能。本方法在这些场景中具有更好的深度预测表现,深度信息更加连贯,物体轮廓清晰,热图像提供了稳定的输入特征,弥补了可见光在弱光下的不足。

此外,在夜间情况下可见光相机通常会降低快门速度,增加传感器的曝光时间以保证图像整体的亮度。但该机制导致画面中的运动物体(如行驶的车辆)出现显著的运动模糊现象,此类运动模糊在现实世界中难以避免,加剧了模型对深度图的错误估计。

相比之下,本方法基于热图像进行深度补全,在视觉效果上提出的方法取得了更合理的结果。所提方法优势在于其在白天与夜间的温度分布相似,热图像在白天和夜间提供了一致的特征输入,深度补全效果不受光照变化的影响。夜间场景中,热成像相机拍摄的图像并未出现严重的运动模糊,且车辆等移动物体的轮廓和三维形状仍然清晰。

2.5　VIVID 数据集上的实验测试

为了进一步验证本文方法在不同数据集下的性能,在 VIVID 数据集上进行了对比评估实验。表 2 展示了提

表 1 MS2 数据集上的深度预测定量对比结果

Table 1 Quantitative comparison results for depth prediction on the MS2 dataset

场景	实验方法	深度图误差评估指标/mm							深度图准确度评估指标		
		MAE	*RMSE*	*IMAE*	*IRMSE*	*Abs Rel*	*Sq Rel*	*RMSE* Log	δ<1.25	δ<1.25^2	δ<1.25^3
白天	VOICED	1 177.11	1 999.84	3.064	5.378	0.050	155.51	0.074	0.981	0.997	0.999
	ScaffNet	683.960	1672.68	2.146	4.865	0.033	169.63	0.063	0.989	0.996	0.998
	FusionNet	649.740	1324.65	2.076	4.281	0.030	69.609	0.051	0.993	0.999	1.000
	KBNet	549.669	1174.86	2.006	4.521	0.027	60.590	0.049	0.993	0.998	1.000
	本文(快速)	414.725	1 090.28	1.210	2.949	0.017	75.888	0.042	0.994	0.997	0.999
	本文	**332.583**	**1 009.97**	**0.968**	**2.501**	**0.015**	**38.334**	**0.034**	**0.996**	**0.999**	**1.000**
多云	VOICED	1 256.08	2 019.61	3.595	6.357	0.057	197.23	0.086	0.970	0.995	0.998
	ScaffNet	704.603	1 701.50	2.472	5.597	0.038	262.05	0.072	0.978	0.992	0.996
	FusionNet	669.693	1 341.74	2.384	5.095	0.034	118.41	0.061	0.983	0.995	0.998
	KBNet	554.413	1 189.66	2.233	5.208	0.030	88.680	0.058	0.985	0.996	0.999
	本文(快速)	314.083	751.804	**1.111**	**2.813**	0.016	29.742	0.033	**0.996**	0.999	1.000
	本文	**301.981**	**748.042**	1.143	2.861	**0.015**	**29.158**	**0.033**	0.994	**0.999**	**1.000**
夜晚	VOICED	945.840	1 578.63	3.174	5.781	0.045	104.69	0.070	0.984	0.997	0.999
	ScaffNet	696.972	1 574.82	2.405	4.923	0.035	138.20	0.065	0.988	0.997	0.998
	FusionNet	675.093	1 270.09	2.374	4.574	0.033	68.315	0.055	0.991	0.999	1.000
	KBNet	608.998	1 195.26	2.501	5.456	0.032	68.557	0.059	0.987	0.997	0.999
	本文(快速)	343.338	**823.654**	1.031	2.290	0.016	30.500	0.033	0.997	0.999	1.000
	本文	**316.101**	859.035	**0.876**	**1.894**	**0.014**	**27.086**	**0.029**	**0.998**	**0.999**	**1.000**

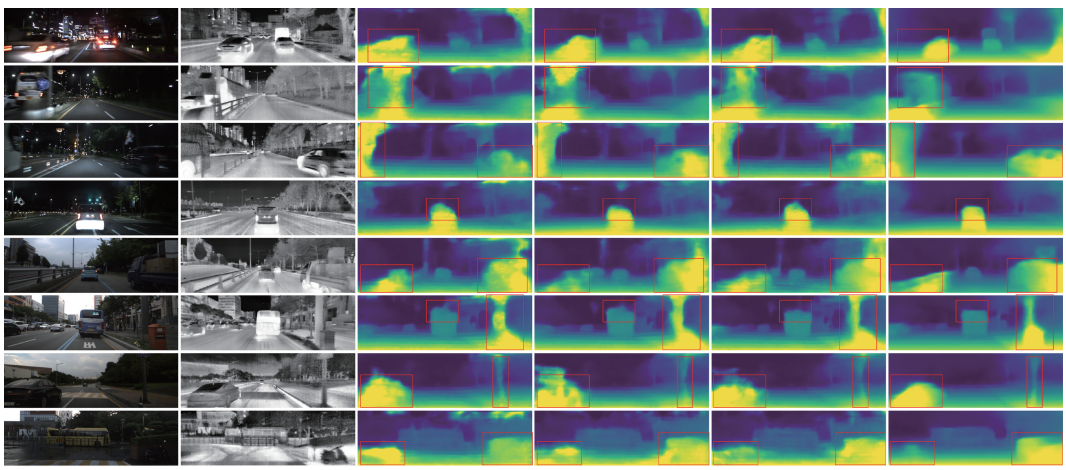


图 7 MS2 数据集上深度预测的视觉定性对比结果

Fig. 7 Visual qualitative comparison results for depth prediction on the MS2 dataset

出的方法与其他方法的定量对比结果,图 10 是各方法的定性视觉效果。在夜间场景下,VIVID 数据集中的场景光照更加昏暗,且点云更为稀疏,在该条件下,定量结果表明所提方法仍然优于其他方法。在 MAE 指标上,提出的方法相比基准方法减少了 266.925 mm,性能提升了 25.28%。

在图像中的完全黑暗的区域,热图像显示出更多的结构和场景细节,为网络模型提供了有效信息。而在这

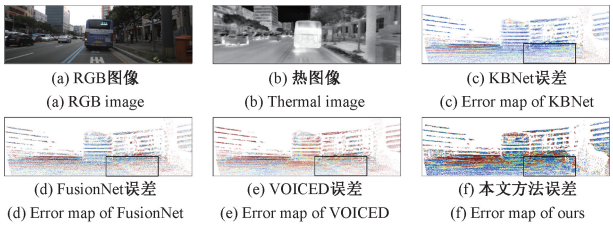


图 8　白天情况 MS2 数据集深度预测误差

Fig. 8　Daytime depth prediction error maps for MS2 dataset

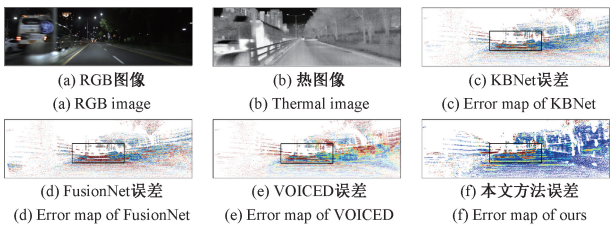


图 9　夜晚情况 MS2 数据集深度预测误差

Fig. 9　Nighttime depth prediction error maps for MS2 dataset

表 2　VIVID 数据集上深度预测定量对比结果

Table 2　Quantitative comparison results for depth prediction on the VIVID dataset

方法	误差/mm				准确度
	MAE	*RMSE*	*Abs Rel*	*Sq Rel*	$\delta<1.25$
VOICED	2 199. 30	3 326. 54	0. 087	397. 68	0. 939
ScaffNet	1143. 48	2 127. 55	0. 052	185. 09	0. 967
FusionNet	1 409. 67	2 513. 94	0. 060	218. 71	0. 959
KBNet	1 056. 02	2 169. 97	0. 043	157. 54	0. 933
本文(快速)	833. 692	1 831. 09	0. 039	125. 56	0. 976
本文	**789. 090**	**1 830. 41**	**0. 036**	**125. 29**	**0. 976**

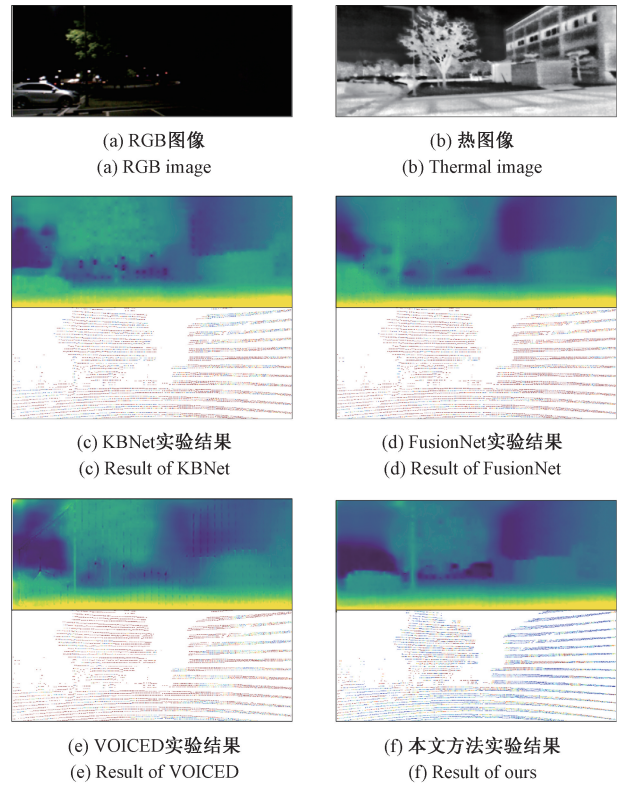


图 10　VIVID 数据集深度预测的视觉结果

Fig. 10　Visual results for depth prediction on the VIVID dataset

些区域,RGB 深度补全方法仅能依赖稀疏深度图进行三维结构学习,难以准确估计出深度。相比之下,本方法利用热图像的优势,能够在光线不足的情况下依然鲁棒地估计稠密深度。

2. 6　消融实验

为了验证提出方法的有效性以及各个子模块对整体系统的贡献,本节在 MS2 数据集上进行了消融实验,实验配置与实施细节与前述部分一致,其定量的实验结果如表 3 和 4 所示。

表 3　各模块效果的深度预测消融实验结果

Table 3　Experimental results of depth predictive ablation for each module effect

方法	深度图误差评估指标/mm							深度图准确度评估指标		
	MAE	*RMSE*	*IMAE*	*IRMSE*	*Abs Rel*	*Sq Rel*	*RMSE* Log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
本文(标准)	**316. 642**	865. 794	**0. 990**	**2. 384**	**0. 015**	**31. 167**	**0. 032**	0. 996	**0. 999**	**1. 000**
本文(去除 MMFF)	358. 347	921. 644	1. 164	2. 750	0. 017	38. 266	0. 036	0. 995	0. 999	1. 000
本文(去除 DTF)	332. 199	**862. 913**	1. 107	2. 505	0. 016	31. 470	0. 032	**0. 997**	0. 999	1. 000
本文(去除注意力)	328. 436	891. 471	1. 019	2. 489	0. 015	35. 549	0. 033	0. 996	0. 999	1. 000
本文(去除热图像重映射)	558. 542	1 369. 78	2. 506	5. 283	0. 030	73. 278	0. 059	0. 982	0. 995	0. 998

1)各模块的效果

通过开启和关闭不同组件的方式,表 3 展示了本方法的各模块效果的消融实验,证明了各组件对所提方法的重要性与贡献。“本文(标准)”表示使用全部组件的结果,除了 RMSE 和 2 个指标之外,模型性能达到了最优,即使在该指标上,模型的性能也仅次于最优结果,

表 4 不同模型版本的深度预测消融实验结果

Table 4 Experimental results of depth predictive ablation with different model versions

方法	深度图误差评估指标/mm							深度图准确度评估指标			耗时	参数	*MACs*
	MAE	*RMSE*	*IMAE*	*IRMSE*	*Abs Rel*	*Sq Rel*	*RMSE* Log	δ<1.25	δ<1.25^2	δ<1.25^3	/ms	/M	/G
本文(标准)	**316.642**	865.794	**0.990**	2.384	**0.015**	**31.167**	**0.032**	**0.996**	**0.999**	**1.000**	21.51	4.00	10.67
本文(快速)	351.064	897.325	1.134	2.683	0.017	36.279	0.035	0.995	0.999	1.000	**15.32**	**1.01**	**2.82**
本文(小型)	361.303	939.246	1.201	2.851	0.017	38.615	0.037	0.995	0.998	0.999	20.53	1.01	2.82
本文(大型)	319.092	**855.965**	1.001	**2.381**	0.015	31.517	0.032	0.996	0.999	1.000	22.50	4.77	19.20

分别仅下降了 0.33% 和 0.07%。在移除不同组件后,模型整体性能表现出不同程度的下降。例如,“本文(去除 MMFF)”的结果表示,去除多模态特征融合模块,并用单个卷积层代替后,MAE 指标下降了 41.71 mm。MMFF 模块作为所提编码器的关键组件,能够引导模型理解不同模态之间的特征距离关系以及进行数据交互。相较于固定尺寸的小卷积网络,MMFF 模块捕捉局部领域的特征关系更为灵活。此外相比单纯的卷积网络仅对所有输入特征施加线性变化,MMFF 模块根据多种模态之间的相似度自适应地调整特征的权重关系,引导不同模态之间更灵活有效的结合。

另一个对模型影响较大的组件是热图像重映射。去除这一模块,直接使用热成像相机提供的原始热图像进行训练学习后,深度图预测的误差的增加较大。“本文(去除热图像重映射)”的结果显示,MAE 指标下降了 76.40 mm。原始热图像的测量范围较宽泛,稀释了热图像中的有效信息,减少了图像之间的差异,影响了模型的学习效果。热图像重映射通过裁切无效的测温范围并调整温度分布,增加了自监督训练框架下相邻帧之间的视差量,有效提升了模型的训练效果。

2)各模型尺寸的效果

为评估提出方法在不同模型尺寸下的性能与推理时间,表 4 展示的不同版本的模型的实验结果对比,证明了“本文(标准)”在模型尺寸与性能之间达到平衡。

各个模型版本的编码器配置如表 5 所示,其中,“[3,32,16,1,8]”表示卷积网络参数为 3×3 核尺寸,热图像分支卷积网络为 32 通道,深度图分支卷积网络为 16 通道,步长为 1,注意力头数为 8。整体而言,提出的模型尺寸相对较小,标准版本的模型参数仅比基准方法中的最小模型 KBNet 增加了 0.97 M,而“本文(快速)”和“本文(小型)”版本的参数量则减少了 3.23 M,但其性能依然优于 KBNet。“本文(快速)”通过简化多模态特征融合模块,降低了自注意力机制的计算复杂度,从而加快了模型的推理时间。具体而言,“本文(快速)”处理每个样本的时间为 15.32 ms(即 65 fps),相比基准版本的推理速度提升了 40.42%,参数量减少了 74.5%,乘法累加运算次数(multiply accumulate operations, MACs)减少了 73.5%。

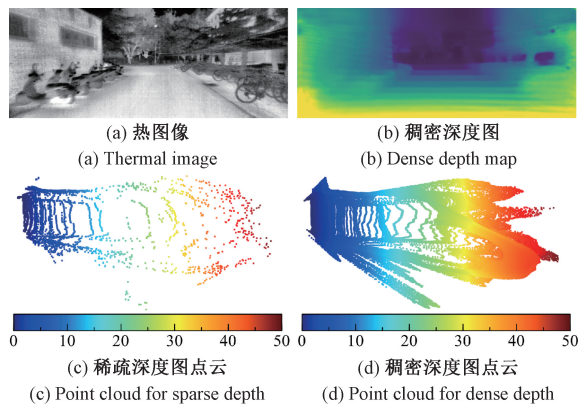
表 5 不同模型版本的编码器配置

Table 5 Encoder configuration for different model versions

网络层级	本文(标准)	本文(小型,快速)	本文(大型)
层级 0	[3,32,16,1,8]	[3,16,8,1,4]	[3,48,16,1,8]
层级 1	[3,32,16,2,8]	[3,16,8,2,4]	[3,48,16,2,8]
层级 2	[3,64,32,2,8]	[3,32,16,2,4]	[3,96,32,2,8]
层级 3	[3,128,64,2,8]	[3,64,32,2,4]	[3,192,64,2,8]
层级 4	[3,256,128,2,8]	[3,128,64,2,4]	[3,256,128,2,8]
层级 5	[3,256,256,2,8]	[3,128,128,2,4]	[3,256,256,2,8]

2.7 实际场景中的实验测试

上述在公开数据集上的对比实验证明了所提方法的性能,在不同光照条件下能够稳定输出稠密的深度图。为验证其在实际场景中的运行效果,网络模型在 MS2 数据集上预训练后,再使用实际场景采集的数据样本进行微调训练,稠密化实验效果如图 11 所示。展示了模型处理前后的深度图点云密度变化。在图 11(c)中点云的密度和分布稀疏,无法全面覆盖整个场景,点云在较远区域的数据几乎缺失。经深度补全后图 11(d)点云密度大幅



(a) 热图像 (a) Thermal image

(b) 稠密深度图 (b) Dense depth map

(c) 稀疏深度图点云 (c) Point cloud for sparse depth

(d) 稠密深度图点云 (d) Point cloud for dense depth

图 11 实际场景中的深度补全效果

Fig. 11 Depth completion result in real scenarios

增加,相比稀疏点云,稠密点云更好地还原了场景的三维结构,在背景和边缘区域点云密度的提升明显,证明了所提的深度补全方法的有效性。

3 结　论

本研究提出了一种基于自监督学习的深度补全方法,从热图像和稀疏深度数据中预测稠密深度图。所提方法探索了热成像相机与激光雷达传感器数据融合方法在全天候深度补全任务中的有效性,并通过层内融合策略和两个模态融合模块,实现了激光雷达与热图像的有效结合。基于自监督的温度重建损失、稀疏深度损失和平滑度损失驱动训练,无需稠密深度图作为真值。实验结果表明,相较于对比方法,所提方法在低光照和无光照条件下能准确稳健地预测深度图。未来工作将进一步探索多传感器融合方法,提高复杂情况下环境感知的鲁棒性和可靠性。

参考文献

[1] 戴虎,郑睿,马小陆, 等. 基于粒子群的多毫米波安防机器人环境感知方法[J]. 仪器仪表学报,2024,41(1):90-100.
DAI H, ZHENG R, MA X L, et al. Environment perception method based on PSO for multi-millimeter wave security robot[J]. Chinese Journal of Scientific Instrument, 2024, 41(1): 90-100.

[2] 余萍,胡旭欣. 基于单目深度估计和校准参数的距离测算方法[J]. 电子测量技术,2022,45(20):88-94.
YU P, HU X X. Distance measurement method based on monocular depth estimation and calibration parameters[J]. Electronic Measurement Technology, 2022, 45(20): 88-94.

[3] 白宇,梁晓玉,安胜彪. 深度学习的 2D~3D 融合深度补全综述[J]. 计算机工程与应用,2023,59(13):17-32.
BAI Y, LIANG X Y, AN SH B. Review of 2D-3D fusion deep completion of deep learning[J]. Computer Engineering and Applications, 2023, 59(13): 17-32.

[4] 周治国,邸顺帆,冯新. 语义信息增强的 3D 激光 SLAM 技术进展[J]. 仪器仪表学报,2023,44(3):209-220.
ZHOU ZH G, DI SH F, FENG X. Advances in SIE 3D LiDAR SLAM technology [J]. Chinese Journal of Scientific Instrument, 2023, 44(3): 209-220.

[5] 赵壮,马国梁. 自适应滤波协同图优化导航方法研究[J]. 仪器仪表学报,2023,44(7):271-281.
ZHAO ZH, MA G L. Research on the adaptive filtering-collaborative graph optimization navigation method[J]. Chinese Journal of Scientific Instrument, 2023, 44(7): 271-281.

[6] 伞红军,杨晓园,陈久朋, 等. 基于拟水流算法在移动机器人路径规划中的应用[J]. 仪器仪表学报,2024,45(7):263-278.
SAN H J, YANG X Y, CHEN J P, et al. Research on path planning of mobile robot based on the stream algorithm[J]. Chinese Journal of Scientific Instrument, 2024, 45(7): 263-278.

[7] 陈慧娴,吴一全,张耀. 基于深度学习的三维点云分析方法研究进展[J]. 仪器仪表学报,2023,44(11):130-158.
CHEN H X, WU Y Q, ZHANG Y. Research progress of 3D point cloud analysis methods based on deep learning[J]. Chinese Journal of Scientific Instrument, 2023, 44(11): 130-158.

[8] HU J J, BAO CH Y, OZAY M, et al. Deep depth completion from extremely sparse data: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8244-8264.

[9] ZHANG Y M, GUO X D, POGGI M, et al. CompletionFormer: Depth completion with convolutions and vision transformers[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18527-18536.

[10] YU ZH, SHENG Z H, ZHOU Z L, et al. Aggregating feature point cloud for depth completion[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 8698-8709.

[11] TANG J, TIAN F P, AN B SH, et al. Bilateral propagation network for depth completion[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9763-9772.

[12] SHI Y X, SINGH M K, CAI H, et al. DeCoTr: Enhancing depth completion with 2D and 3D attentions[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 10736-10746.

[13] YAN ZH Q, LIN Y K, WANG K, et al. Tri-perspective

view decomposition for geometry-aware depth completion[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 4874-4884.

[14] WONG A, CICEK S, SOATTO S. Learning topology from synthetic data for unsupervised depth completion[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1495-1502.

[15] WONG A, FEI X H, TSUEI S, et al. Unsupervised depth completion from visual inertial odometry[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 1899-1906.

[16] ZHANG Q, CHEN X Y, WANG X G, et al. Self-supervised depth completion based on multi-modal spatio-temporal consistency [J]. Remote Sensing, 2023, 15(1): 135-151.

[17] JEON J, LIM H, SEO D V, et al. Struct-MDC: Mesh-refined unsupervised depth completion leveraging structural regularities from visual SLAM [J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6391-6398.

[18] YAN ZH Q, WANG K, LI X, et al. DesNet: Decomposed scale-consistent network for unsupervised depth completion [C]. 37th AAAI Conference on Artificial Intelligence, 2023: 3109-3117.

[19] LI T, WU D D, ZHOU M H, et al. ADCV: Unsupervised depth completion employing adaptive depth-based cost volume[J]. Digital Signal Processing, 2024, 155(1): 104750-104762.

[20] WONG A, SOATTO S. Unsupervised depth completion with calibrated backprojection layers[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 12727-12736.

[21] 潘冬,李奕天,马晓路, 等. 粉尘干扰下工业物料表面单视角三维热成像方法[J]. 仪器仪表学报,2024, 45(10):143-153.
PAN D, LI Y T, MA X L, et al. Single-view 3D thermography method for industrial material surfaces under dust interference[J]. Chinese Journal of Scientific Instrument, 2024, 45(10): 143-153.

[22] ZHOU T H, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6612-6619.

[23] WANG ZH, 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.

[24] GODARD C, AODHA O M, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition,2017:6602-6611.

[25] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[26] SHIN U, PARK J, KWEON I S. Deep depth estimation from thermal image[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 1043-1053.

[27] LEE A J, CHO Y, SHIN Y S, et al. VIVID++: Vision for visibility dataset[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6282-6289.

[28] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[29] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network [C]. Proceedings of the 28th International Conference on Neural Information Processing Systems, 2014, 2: 2366-2374.

[30] GAIDON A, WANG Q, CABON Y, et al. Virtual worlds as proxy for multi-object tracking analysis [J]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016:4340-4349.

作者简介

于睿,2020 年于南京理工大学获得学士学位,现为南京理工大学博士研究生,主要研究方向为深度学习机器视觉。
E-mail:yu831@foxmail.com

Yu Rui received his B. Sc. degree from Nanjing University of Science and Technology in 2020. He is currently a Ph. D. candidate at Nanjing University of Science and Technology. His main research interest is deep learning machine vision.

马国梁(通信作者),2006 年于南京理工大学获得博士学位,现为南京理工大学副教授,主要研究方向为机器人系统与控制。
E-mail:mgl@ njust. edu. cn

Ma Guoliang (Corresponding author) received his Ph. D. degree from Nanjing University of Science and Technology in 2006. He is currently an associate professor at Nanjing University of Science and Technology. His main research interests include robot systems and control.

郭健,2002 年于南京理工大学获得博士学位,现为南京理工大学教授、博士生导师,主要研究方向为智能控制与智能系统。
E-mail:guoj1002@ njust. edu. cn

Guo Jian received his Ph. D. degree from Nanjing University of Science and Technology in 2002. He is currently a professor and a Ph. D. advisor at Nanjing University of Science and Technology. His main research interests include intelligent control and intelligent systems.

许立松,2018 年于南京理工大学获得学士学位,现为南京理工大学博士研究生,主要研究方向为机器人形态感知技术。
E-mail:xulisong@ njust. edu. cn

Xu Lisong received his B. Sc. degree from Nanjing University of Science and Technology in 2018. He is currently a Ph. D. candidate at Nanjing University of Science and Technology. His main research interest is robot form perception technology.