DiffusionMVS:基于扩散约束的遥感影像立体重建算法

连远锋[1,2]，王森[1]

1. 中国石油大学(北京) 人工智能学院, 北京 102249;
2. 中国石油大学(北京)石油数据挖掘北京市重点实验室, 北京 102249

摘 要：针对遥感影像多视图立体任务中存在的特征匹配精度低、预测深度图存在噪声和边缘重建不完整等问题，本文提出一种基于扩散约束的多视图立体网络DiffusionMVS（Diffusion Multi-View Stereo）。首先，在特征金字塔网络的基础上，设计基于特征增强的多尺度特征提取模块MFE-FPN（Multi-scale Feature Enhancement Feature Pyramid Network）来增强网络学习多视图遥感影像特征的能力；其次，提出自适应特征聚合模块AFA（Adaptive Feature Aggregation）来动态整合不同层次的特征以捕获目标边缘的深度细节特征；最后，设计基于扩散约束的代价体优化模型DCM（Diffusion Constrained Module），通过优化存在噪声点的深度值分布来消除预测深度图存在的噪声干扰，并结合边缘引导的Transformer网络优化深度图边缘重建效果。实验结果显示，在WHU-TLC和LuoJia-MVS数据集测试中，与基准模型相比，本文提出的模型DiffusionMVS网络的平均绝对误差（Mean Absolute Error，MAE）指标分别提升了28.11%和3.37%，展示了较好的重建性能和泛化能力。

关键词：遥感图像，多视图立体，多尺度特征提取，自适应特征聚合，扩散模型，边缘引导Transformer

中图分类号：TP701/P2

引用格式：连远锋,王森.2026.DiffusionMVS:基于扩散约束的遥感影像立体重建算法.遥感学报,30(5):1510-1523

Lian Y F and Wang S. 2026. DiffusionMVS: Multi-view stereo reconstruction algorithm for remote sensing image based on diffusion constraints. National Remote Sensing Bulletin, 30(5):1510-1523[DOI:10.11834/jrs.20265119]

1 引 言

基于遥感影像的大规模场景三维重建对智慧城市构建、地图导航、虚拟现实、数字孪生等领域具有重要意义（Luo等，2024）。现有的三维重建算法通常依赖于特征匹配技术（吴博剑和黄惠，2020），在小规模或简单场景中表现良好，但在复杂或大规模的场景下，由于存在地表复杂信号及噪声干扰，在目标边缘重建精度和完整性上表现不佳（Zhou等，2021）。近年来，深度学习方法被广泛应用于多视图立体MVS（Multi-View Stereo）重建任务（Liu等，2023a；Gao等，2023；Mao等，2024）。由于成像传感器对环境敏感而造成的噪声干扰会影响图像的特征提取和匹配过程，这会导致深度估计误差增大，进而降低重建场景精度（Han等，2022）。如何降低场景噪声干扰并提升深度图边缘重建效果仍是亟待解决的难题。

多视图立体重建MVS Reconstruction（Multi-View Stereo Recstruction）旨在通过从相机参数已知的多视角图像中恢复场景的3D几何结构（鄢化彪等，2023）。传统的MVS方法大多通过多视图之间的投影关系来计算每个像素点的深度信息。Merrell等（2007）使用平面扫描算法（Plane-Sweeping）（Gallup等，2007）计算目标场景的深度信息，提出基于像素可视性的深度图融合算法。Barnes等（2009）基于随机搜索和迭代优化快速定位，提出Patchmatch算法提高了三维重建处理速度和匹配的准确性。Schönberger和Frahm（2016）提出COLMAP方案，利用光度一致性估计视角的深度值和法向量值，提高重建场景的完整性。Xu和Tao（2019）提出ACMM算法，通过采用多尺度块匹配、自适应棋盘格采样和多假设联合的视图选

收稿日期：2025-04-03；**预印本**：2025-05-09

基金项目：国家自然科学基金(编号:61972353);中国石油天然气集团有限公司-中国石油大学(北京)战略合作科技专项(编号:ZLZX2020-05)

第一作者简介：连远锋,主要研究方向为遥感图像处理。E-mail：lianyuanfeng@cup.edu.cn

择方案，提升深度估计的精度。尽管传统的MVS重建方法取得了很大进步，但在处理遮挡、纹理缺乏或光照变化大的复杂场景时，仍面临重建精度低和完整性不佳的问题。

近年来，基于深度卷积神经网络DCNN（Deep Convolutional Neural Networks）的多视图立体重建得到了广泛应用。Yao等（2018）将卷积神经网络与MVS相结合，提出基于深度学习的MVS网络（MVSNet），通过使用CNN提取图像特征，利用可微单应变换构造代价体（Cost Volume）并使用3D U-Net对代价体进行正则化，从而实现多视图深度估计。基于此，Yao等（2019）针对内存消耗问题提出基于循环神经网络的R-MVSNet网络，通过将MVSNet中的3D CNN模块替换为门控GRU（Gate Recurrent Unit）单元，减少了内存消耗。Yu和Gao（2020）提出Fast-MVSNet，通过稀疏代价体推断初始稀疏深度图来缩短运行时间。Gu等（2020）提出级联结构的Cas-MVSNet网络，采用图像特征金字塔并通过级联策略以粗到细的方式来估计深度图。Cheng等（2020）在级联架构基础之上提出UCS-Net，通过构建自适应代价体来提高深度图的分辨率和精度。Wei等（2021）针对多尺度上下文信息缺失问题，提出一种混合递归正则化网络Aa-RMVSNet，通过聚合多尺度上下文信息高效处理原始大小的代价体。Zhang等（2023a）基于3D UNet设计注意力形状感知网络DSC-MVSNet用于提高模型重建效率。Zhang等（2023b）提出ARAI-MVSNet自适应地进行全像素深度范围预测及深度间隔划分，从而生成准确的深度图。

尽管基于深度学习的多视图立体重建方法在室内场景取得较大进展，然而在大规模场景重建过程中，仍然存在内存消耗大、重建精度低等问题。为此，Liu和Ji（2020）提出遥感影像的多视图立体网络（RED-Net），通过采用递归编码器—解码器RED（Recurrent Encoder-Decoder）结构实现代价体的顺序正则化，在重建效率和精度方面均有显著提升。基于此，Yu等（2021）针对复杂建筑重建不佳问题提出MS-REDNet网络，通过将语义分割和MVS方法结合提升建筑重建的完整性。Gao等（2021）通过将有理多项式相机RPC（Rational Polynomial Camera）模型融入MVS框架来校正遥感图像中的非线性畸变，提出Sat-MVS来增强图像间的对应关系，进而提升重建精度。Lin等（2023）针对匹配空洞问题提出A-SATMVSNet网络，通过三重膨胀卷积和注意力机制优化特征提取。Li等（2023）提出分层可变形的级联MVS网络结构（HDC-MVSNet），通过全尺度特征提取和分层代价体构建模块来处理大尺寸的遥感图像。Zhang等（2024）针对边缘特征缺失问题，提出EG-MVSNet网络，通过整合边缘信息到MVS网络中来细化建筑物深度估计精度。综上所述，目前基于遥感图像的立体重建已有诸多成果。然而，由于遥感图像覆盖范围广、分辨率高、地物背景复杂和大气散射等因素导致遥感图像存在高斯噪声，这干扰了图像中特征提取和匹配过程的计算结果，进而导致深度估计精度降低。

扩散模型（Ho等，2020）是一种基于马尔可夫链的生成模型，通过迭代去噪过程将来自标准高斯分布的样本转换为来自经验数据分布的样本。扩散模型在遥感图像处理任务中展示了出色的性能，例如遥感图像变化检测（Wen等，2024）、语义分割（Toker等，2024）、目标检测（Wang等，2024）和超分辨率重建（Dong等，2024）等。Khan等（2021）通过优化稀疏点集来估计密集深度图，利用可微扩散约束最小化多视图重投影误差来提高复杂场景的深度估计精度。Shao等（2022）提出Diffustereo残差图扩散模型，通过迭代细化视差图实现精准的人体三维重建。Heo和Lee（2024）利用去噪扩散模型对深度图进行校正和优化。Li等（2024）提出Sat2Scene网络通过扩散模型来生成纹理颜色并基于神经渲染技术实现场景渲染。

因此，为提高遥感图像重建精度，本文提出一种基于扩散约束的MVS网络（DiffusionMVS）。首先，设计基于特征增强的多尺度特征提取模块来提高遥感影像多尺度特征表达能力；其次，提出自适应特征聚合模块来实现不同特征层级的动态聚合，进一步提升对目标边缘特征的识别能力；最后，为消除预测深度图中存在的噪声干扰，提出基于扩散约束的深度图优化调整模型，通过边缘引导的Transformer模型来提升深度图的边缘重建效果，以期为后续高精度三维重建提供边缘细节丰富的深度图。

2 研究方法与构建原理

针对遥感影像多视图立体中预测深度图存在的噪声干扰问题，本文提出基于扩散约束的遥感

影像多视图立体网络（DiffusionMVS），整体框架如图1所示。该模型主要由多尺度特征增强提取模块MFE-FPN（Multi-scale Feature Enhancement Feature Pyramid Network）、自适应特征聚合模块AFA（Adaptive Feature Aggregation）和扩散约束模块DCM（Diffusion Constrained Module）3部分构成。DiffusionMVS网络输入N张多视图的遥感图像，其中第一张为参考图像$\boldsymbol{I}_1$，其余N-1张为源图像$\boldsymbol{I}_{N-1}$，采用由粗到细的策略通过三阶段逐级预测深度图。该网络首先通过MFE-FPN模块对遥感图像进行多尺度特征提取，生成多尺度特征图。FPN顶层特征图经过边缘感知网络映射计算，得到边缘感知特征并用于多尺度特征融合。自适应特征聚合模块对多尺度特征进行聚合，进而形成匹配代价体特征。扩散约束模块将匹配代价体特征和边缘感知特征融合，通过边缘引导Transformer模块来提高去噪阶段的边缘细节特征表达能力。最后对代价体特征进行正则化和深度回归运算，生成深度图重建结果。

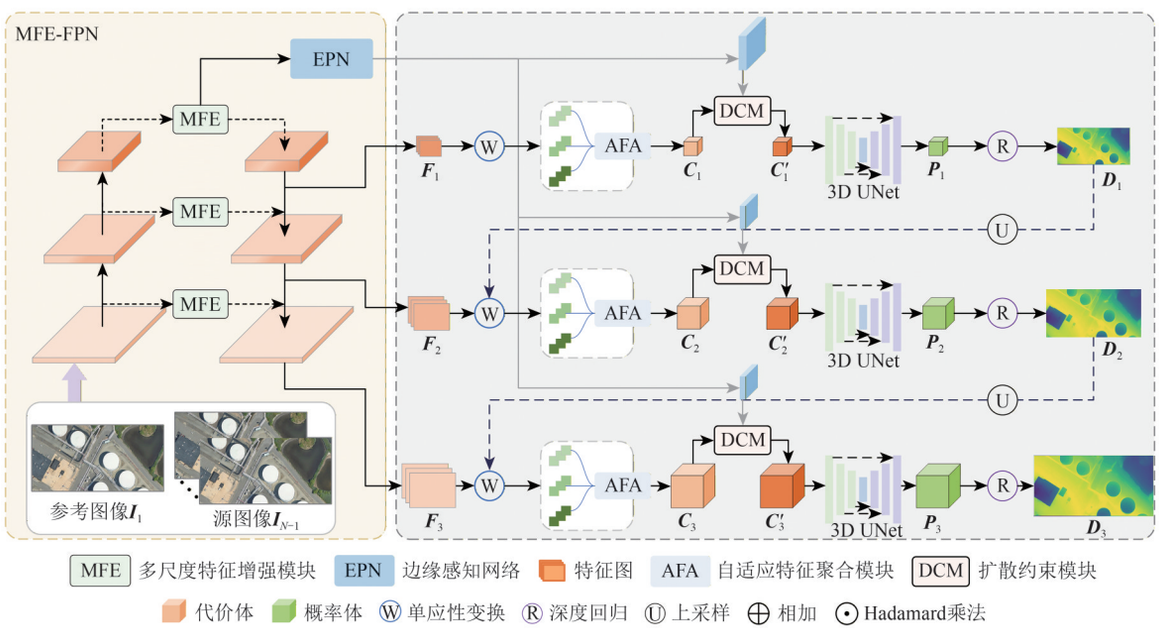


图1 整体网络结构

Fig. 1 The overall structure

2.1 多尺度特征增强模块

为了增强网络对多视图遥感影像的特征表达能力，本文基于动量梯度下降算法的优化多尺度特征增强模块MFE-FPN，具体推导过程如下：

$$x_{k+1}=x_k-\alpha\nabla f(x_k)-\beta\nabla f(x_{k-1}) \tag{1}$$

式中，α和β分别为动量参数和学习率，$\nabla f(x_k)$表示目标函数f在x_k处的梯度，f为满足Lipschitz性质的连续函数。优化α和β的组合可加速网络收敛。

设神经网络在传播过程中，梯度的更新格式为

$$x_{k+1}=\sum_{i=0}^{k}\alpha_i^{k+1}x_i+\sum_{i=0}^{k+1}\beta_i^{k+1}\nabla f(x_i) \tag{2}$$

式中，α^k、β^k表示第k步更新的步长。

将梯度项$\nabla f(x_i)$替换为模块$\boldsymbol{T}$，得到L层的神经网络设计架构为

$$\begin{aligned}&z_{k+1}=\sum_{i=0}^{k}\alpha_i^{k+1}z_i+\sum_{i=0}^{k+1}\beta_i^{k+1}\boldsymbol{T}_i^{k+1}(z_i),k=1,2,\cdots,L-1\\&z_{L+1}=\boldsymbol{W}^{L+1}z_L+\boldsymbol{b}^{L+1}\end{aligned} \tag{3}$$

式中，z_k为网络中第k层的输出，$\boldsymbol{T}$为包含两层网络结构的模块，定义为

$$\boldsymbol{T}_i^k(z)=\boldsymbol{V}_i^k\alpha\left(\boldsymbol{W}_i^kz+\boldsymbol{b}_i^k\right) \tag{4}$$

式中，$\boldsymbol{V}^k$、$\boldsymbol{W}^k$是权重矩阵，$\boldsymbol{b}^k$是权重系数，α为满足特定条件的激活函数。

进一步，将上述模块$\boldsymbol{T}$替换为双层卷积结构（Wu等，2024），可以得到：

$$\boldsymbol{T}_j^k(z_j)=\boldsymbol{V}_j^k\sigma_{\mathrm{R}}\left(\boldsymbol{W}_j^k\sigma_{\mathrm{R}}(z)+\boldsymbol{b}_{j,1}^k\right)+\boldsymbol{b}_{j,2}^k \tag{5}$$

式中，σ_{R}表示ReLU激活函数，$\boldsymbol{b}_{j,1}^k$、$\boldsymbol{b}_{j,2}^k$分别表示第j层的偏置系数。

根据式（1）和（2），得到推导的神经网络可表示为

$$z_{k+1} = z_k + \alpha \boldsymbol{T}^k(z_k) + \beta(z_k - z_{k-1}) \tag{6}$$

MFE模块如图2所示，在MFE模块中加入残差连接，通过将浅层的细节特征与高层的语义特征相融合，提升了网络对多视图遥感图像特征提取能力。MFE-FPN模块由编码器和带跳连接的解码器组成。该模块输出一个三尺度的特征金字塔，其大小分别为输入遥感图像大小的｛1/16，1/4，1｝，特征通道数分别为32、16和8，输入图像通过多个卷积模块生成特征图$(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3)$，通过MFE模块自上而下逐级融合得到更为精细的特征图$(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3)$。

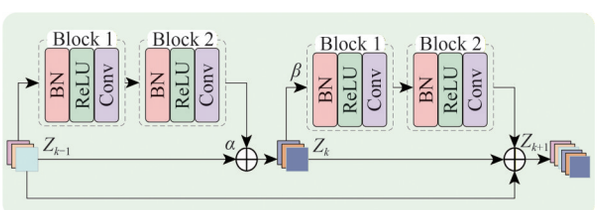


图2 MFE模块

Fig. 2 MFE module

如图3所示，边缘感知网络EPN（Edge Perception Network）以高层特征作为输入，首先对输入的特征通过包含3×3卷积、ReLU激活函数和1×1卷积的解码器模块进行预处理；然后，通过上采样操作逐步恢复空间细节。为了提高边缘定位精度，上采样后的高层特征与底层特征进行跨尺度融合，实现语义引导。最终，融合后的特征经3×3卷积和Sigmoid激活函数得到边缘特征。

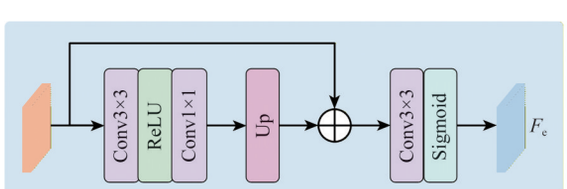


图3 EPN模块

Fig. 3 EPN module

2.2 自适应特征聚合模块

基于深度学习的MVS网络通常使用平面扫描法（Gallup等，2007）构建代价体，通过对深度值进行几何投影实现特征图从源视图平面到参考视图平面的变换。假设参考图像为$\boldsymbol{I}_1$，源图像为$\boldsymbol{I}_i$，利用单应性变换将第i个视角图像的特征图$\boldsymbol{F}_i$映射投影到参考图像$\boldsymbol{I}_1$所对应的平行平面上，则有：

$$\boldsymbol{H}_i(d) = \boldsymbol{K}_i \cdot \boldsymbol{R}_i\left(\boldsymbol{I} - \frac{(\boldsymbol{t}_1 - \boldsymbol{t}_i)\boldsymbol{n}_1^{\mathrm{T}}}{d}\right)\boldsymbol{R}_1^{\mathrm{T}} \cdot \boldsymbol{K}_1^{\mathrm{T}} \tag{7}$$

式中，$\boldsymbol{H}_i(d)$表示在深度d处第i个特征图$\boldsymbol{F}_i$与参考特征图的单应矩阵，参数$\boldsymbol{K}_i$、$\boldsymbol{R}_i$和$\boldsymbol{t}_i$分别为源图像的相机内参矩阵与相机旋转矩阵和平移向量，$\boldsymbol{I}$是单位矩阵，$\boldsymbol{n}_1^{\mathrm{T}}$是参考图像平面法向量的转置。

特征图在通过转换投影到参考视图后，得到特征体$\{\boldsymbol{V}_i\}_{i=1}^{N}$。由于不同特征体对特征匹配的贡献不同，为更好地聚合来自不同尺度的深度特征，减小匹配误差的影响，本文提出一种新的自适应特征聚合模块AFA，如图4所示。AFA模块对输入的特征体$\{\boldsymbol{V}_i\}_{i=1}^{3}$采用不同尺度的编码器来分别处理多尺度信息，对于包含最多信息的低分辨率特征，使用最大尺度的编码器，而低分辨率特征由于信息量相对较少，使用较小尺度的编码器。并使用多层感知机（MLP）进行特征缩放和偏移调整，将特征沿特征通道方向重新进行组合缩放，来实现特征在语义空间中的对齐，从而放大细节深度特征，抑制无关噪声，实现像素级的自适应聚合，即：

$$\boldsymbol{V}_i^{\mathrm{out}} = \left(1 + M_{\alpha i}(\boldsymbol{v}_i)\right) \odot \mathrm{BN}\left(\boldsymbol{V}_i^{\mathrm{in}}\right) \oplus M_{\beta i}(\boldsymbol{v}_i) \tag{8}$$

式中，$\boldsymbol{v}_i$，$i \in \{1, 2, 3\}$表示输入的特征体，BN表示批量归一化，$M_{\alpha i}$和$M_{\beta i}$表示多层感知机，α表示沿通道方向进行特征缩放，β表示特征偏移量。

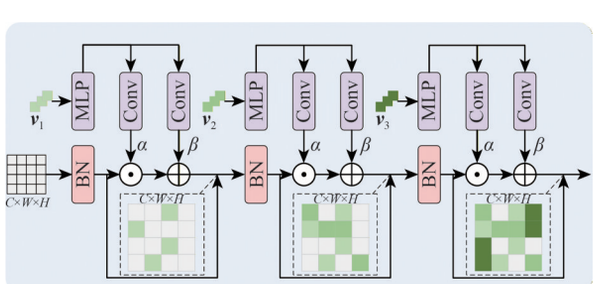


图4 AFA模块

Fig. 4 AFA module

最后，基于方差的成本度量CM（Cost Metric）来衡量匹配相似度，将特征体聚合为代价体$\boldsymbol{C}$，即：

$$\boldsymbol{C} = \frac{\sum_{i=1}^{N}\left(\boldsymbol{V}_i - \bar{\boldsymbol{V}}\right)^2}{N} \tag{9}$$

式中，$\boldsymbol{V}_i$表示输入特征体，$\bar{\boldsymbol{V}}$表示特征体的平均值，N表示输入视图数。

为了从代价体中得到深度图，首先使用3D U-net正则化网络得到概率体（Probability Volume）

$\boldsymbol{P}$，即：

$$\boldsymbol{P}(d) = \sigma(-c_d) \tag{10}$$

式中，$\boldsymbol{P}(d)$表示像素点深度为d的概率，c_d表示在深度为d处的代价体，σ为Softmax函数。

将所有深度假设值与对应的概率估计加权回归得到深度图$\boldsymbol{D}$，即：

$$\boldsymbol{D} = \sum_{d=d_{\min}}^{d_{\max}} d \times \boldsymbol{P}(d) \tag{11}$$

式中，$d_{\min}$表示最小深度值，$d_{\max}$表示最大深度值，d表示预先定义的深度采样值。

2.3 扩散约束模块

扩散约束模块结构如图5所示，在扩散阶段，首先通过学习真实深度图的深度像素分布，对初始深度图的每个像素点进行离散化得到对应的深度值分布$\boldsymbol{X}_0$，即像素级深度概率分布，作为扩散模型的初始化先验样本。之后不断地添加高斯噪声，使其逐步满足高斯分布，即：

$$\boldsymbol{X}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\cdot\epsilon, t = 1,2,\cdots,T \tag{12}$$

式中，$\boldsymbol{X}_t$表示t时刻的深度值分布，$\bar{\alpha}_t$表示噪声系数的平均值，ϵ表示噪声。

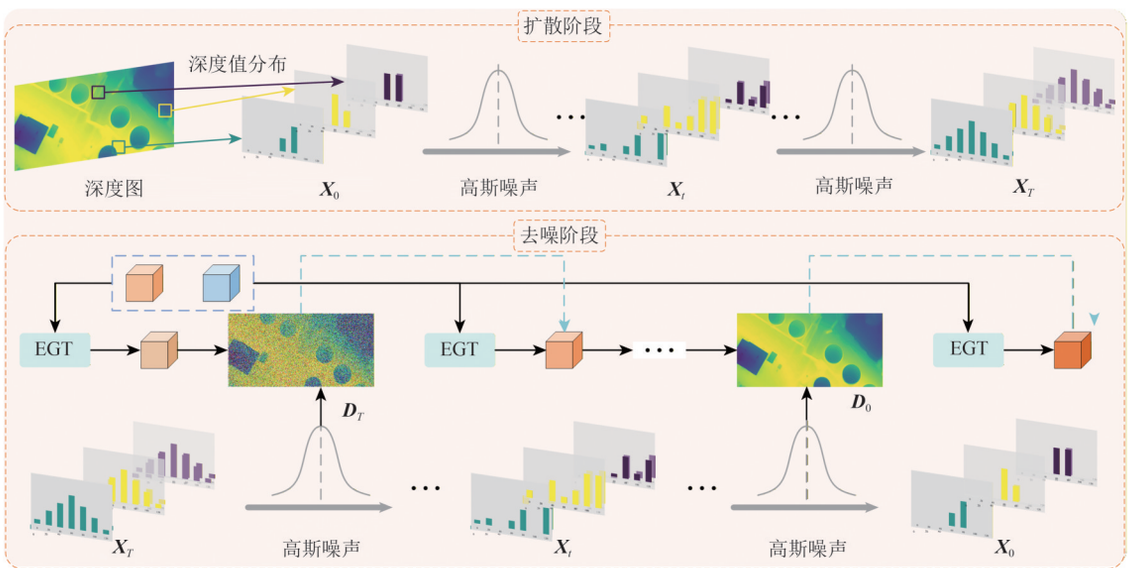


图5 DCM模块

Fig. 5 DCM module

为了提升深度图去噪过程中边缘重建精度，本文设计了基于边缘引导Transformer模块EGT（Edge-guided Transformer），通过在扩散约束模块的迭代去噪阶段实现对深度图边缘引导，如图6所示。对于输入的多尺度代价体特征$\boldsymbol{F}$和边缘感知特征$\boldsymbol{F}_e$，首先，对边缘感知特征上采样到与多尺度代价体特征相同的尺寸，得到$\boldsymbol{F}'_e$，接着，通过空间注意力模块（Hu等，2018）和通道注意力模块（Woo等，2018）计算权重α_{ch}与α_{sp}，分别捕捉空间和通道维度上的关键特征分布，抑制无关信息的干扰。然后，对输入特征与生成的权重α_{ch}与α_{sp}通过乘法与加法操作，得到特征$\boldsymbol{F}_{ch}$和$\boldsymbol{F}_{sp}$，并进行拼接。最后，与$\boldsymbol{F}$和$\boldsymbol{F}_e$连接融合得到最终的多尺度边缘特征$\hat{\boldsymbol{F}}$，从而提升网络对复杂环境中边缘区域的识别和解析能力，该过程可以表示为

$$\begin{cases} \boldsymbol{F}_u = \mathrm{Up}(\boldsymbol{F}_e) \\ \boldsymbol{F}_c = \mathrm{Conv}(\boldsymbol{F}) \\ \boldsymbol{F}'_e = \mathrm{Up}\big(\mathrm{Conv}\big((\boldsymbol{F}_e)\big)\big) \end{cases} \tag{13}$$

$$\begin{cases} \alpha_{sp} = \mathrm{SAM}(\boldsymbol{F}_c, \boldsymbol{F}_u) \\ \alpha_{ch} = \mathrm{CAM}(\boldsymbol{F}_c, \boldsymbol{F}_u) \end{cases} \tag{14}$$

$$\begin{cases} \boldsymbol{F}_{ch} = \alpha_{sp}\cdot\boldsymbol{F}_u + (1-\alpha_{ch})\cdot\boldsymbol{F}_c \\ \boldsymbol{F}_{sp} = \alpha_{ch}\cdot\boldsymbol{F}_u + (1-\alpha_{sp})\cdot\boldsymbol{F}_c \\ \hat{\boldsymbol{F}} = \mathrm{Conv}\big(\mathrm{Concat}(\boldsymbol{F}, \boldsymbol{F}_{cs}, \boldsymbol{F}'_e)\big) \end{cases} \tag{15}$$

式中，CAM表示通道注意力模块，SAM表示空间注意力模块。

去噪阶段将MVS各分支生成的边缘多尺度特征$\hat{\boldsymbol{F}}$和扩散阶段生成的满足高斯分布的深度值$\boldsymbol{X}_t$作为输入，迭代过滤代价体中的噪声信息，得到细化的代价体$\boldsymbol{C}'$，如图5去噪阶段所示。其中，扩散阶段添加的噪声，计算公式如下：

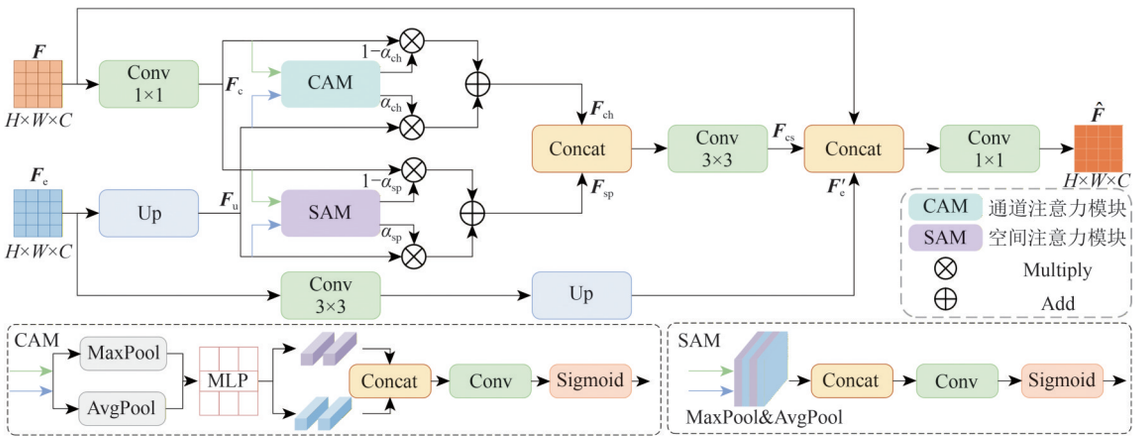


图6 基于边缘引导的Transformer模块

Fig. 6 Edge-guided transformer module

$$\epsilon = \frac{1}{\sqrt{1-\bar{\alpha}_T}}\left(\boldsymbol{X}_T - \sqrt{\bar{\alpha}_T}\,\boldsymbol{X}_0\right) \tag{16}$$

进一步，使用预训练的DDIM（Song等，2022）采样策略恢复$T-1$时刻的深度分布$\boldsymbol{X}_{T-1}$，即：

$$\begin{aligned}\boldsymbol{X}_{T-1} &= \sqrt{\bar{\alpha}_{T-1}}\,\boldsymbol{X}_T + \sqrt{1-\bar{\alpha}_{T-1}-\sigma^2\epsilon} + \sigma\epsilon^* \\ \sigma &= \eta\sqrt{\left(1-\frac{\bar{\alpha}_T}{\bar{\alpha}_{T-1}}\right)\cdot\frac{1-\bar{\alpha}_{T-1}}{1-\bar{\alpha}_T}}\end{aligned} \tag{17}$$

式中，η表示采样置信度，ϵ^*表示高斯噪声。

去噪阶段利用边缘信息引导，减少深度图中的噪声，生成更加精确和平滑的深度图，即：

$$\boldsymbol{D}_t = \mathrm{ReverseDiffusion}\left(\boldsymbol{D}_T, f(\boldsymbol{X}_t, \hat{\boldsymbol{F}})\right) \tag{18}$$

$$f(\boldsymbol{X}_t, \hat{\boldsymbol{F}}) = \mathrm{U-Net}_\theta\left(\mathrm{Concat}(\boldsymbol{X}_t, \hat{\boldsymbol{F}})\right) \tag{19}$$

式中，$\boldsymbol{D}_t$为t时刻的深度图，f表示拟合函数，θ表示网络权重，$\boldsymbol{X}_t$为噪声样本，$\hat{\boldsymbol{F}}$为条件输入特征。

最后，得到去噪阶段t时刻的代价体$\boldsymbol{C}'_t$，即：

$$\boldsymbol{C}'_t = \Gamma(\boldsymbol{D}_t)\odot\left(\mathrm{Conv3D}(\hat{\boldsymbol{F}}) + \mathrm{MLP}(t)\right) \tag{20}$$

式中，Γ表示几何反投影，Conv3D表示3D卷积，MLP表示多层感知机。

2.4 损失函数

本文采用有监督方式训练DiffusionMVS网络，并使用L_1损失来度量DiffusionMVS各阶段的真实深度与预测深度之间的绝对误差，总损失定义为三阶段损失的加权和，即：

$$L_{\mathrm{DE}} = \sum_{k=1}^{3} w_k \sum_{p\in\Omega}^{N} \left\| D^k_{\mathrm{GT}}(p) - \hat{D}^k(p) \right\|_1 \tag{21}$$

式中，L_{DE}表示真实深度图和预测深度图之间的绝对平方差，w_k代表第k阶段的损失权重，Ω表示参考视图中深度有效的像素，$D_{\mathrm{GT}}(p)$是在像素点p处的真实深度值，$\hat{D}(p)$为在同一像素点处的预测深度值。

为减少DiffusionMVS预测深度图时的边缘信息丢失，引入边缘损失函数（Ibrahimli等，2023）来增强模型在深度图边缘区域预测的精度，使用Sobel和Laplacian边缘检测算子进行两阶段边缘深度值提取。边缘损失函数定义为

$$L_{\mathrm{ED}} = \frac{1}{N}\sum_{p\in\Omega}\ell_2\left(E(p), M\left(\varsigma\left(\Delta\widehat{D}(p), \xi\right)\cap\phi\left(\Delta\widehat{D}(p), \tau\right)\right)\right) \tag{22}$$

式中，$E(p)$表示网络在像素点p处预测的边缘深度值，M表示人工标记的边缘掩模，ς和ϕ分别表示Sobel和Laplacian算子结果的阈值函数，ξ和τ表示阈值。最终网络的整体训练损失函数为上述各项损失函数之和：

$$L_{\mathrm{total}} = \lambda_{\mathrm{DE}} L_{\mathrm{DE}} + \lambda_{\mathrm{ED}} L_{\mathrm{ED}} \tag{23}$$

式中，λ_{DE}和λ_{ED}为超参数。

3 实验设计与结果分析

3.1 数据集

为了验证本文提出方法的有效性，使用WHU-TLC（Gao等，2021）、LuoJia-MVS（Li等，2023）两个公开的遥感三维重建数据集以及自建的油气站场多视图数据集进行实验。WHU-TLC为光学卫星影像数据集，经过裁剪处理后划分为6802组768×384大小的图像。其中5011组用于训练，1791组用于测试。

LuoJia-MVS为航空影像数据集，由5680组五视图图像组成，并附有像素级深度图和准确的相机参数，数据集中的每张图像的大小784×368，空间分辨率为10 cm。数据集以约3∶1的比例分为训练集和测试集，其中4320组用于训练，1360组用于测试。

自建的油气站场多视图数据集由无人机拍摄，经过裁剪处理后划分为2000组768×384像素大小的图像，主要用于模型的泛化性测试。

3.2 评价指标

本文主要采用平均绝对误差（MAE）作为评价指标，考虑到WHU-TLC和LuoJia-MVS数据集的场景深度范围不同，因此采用额外的评价指标对模型进行评估。对于WHU-TLC数据集，采用均方根误差（RMSE）、有效网格占比$PAG_{2.5\,m}$和$PAG_{7.5\,m}$、完整性（Comp），综合评估重建的深度图质量。对于LuoJia-MVS数据集，采用$PAG_{0.6\,m}$和误差小于3个深度间隔的像素百分比（<3-interval）进行评估。

MAE表示地面实况和估计深度图之间所有网格单元的L_1距离的平均值。计算公式为

$$\mathrm{MAE}=\frac{\sum_{(i,j)\in \boldsymbol{G}\cap\hat{\boldsymbol{G}}}\left|h_{ij}-\hat{h}_{ij}\right|}{\sum_{(i,j)\in \boldsymbol{G}\cap\hat{\boldsymbol{G}}} I\left((i,j)\in \boldsymbol{G}\cap\hat{\boldsymbol{G}}\right)} \tag{24}$$

式中，$\boldsymbol{G}$和$\hat{\boldsymbol{G}}$分别代表预测深度图和真实深度图的有效网格单元，h_{ij}和$\hat{h}_{ij}$分别指第i行j列像素单元中的预测深度和真实深度，$I(A)$表示当A为真时，值为1；否则值为0。

RMSE表示深度预测值和地面真实值之间的标准差。计算公式为

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{(i,j)\in \boldsymbol{G}\cap\hat{\boldsymbol{G}}}(h_{ij}-\hat{h}_{ij})^2}{\sum_{(i,j)\in \boldsymbol{G}\cap\hat{\boldsymbol{G}}} I\left((i,j)\in G\cap\hat{G}\right)}} \tag{25}$$

$PAG_{2.5\,m}$、$PAG_{7.5\,m}$和$PAG_{0.6\,m}$表示L_1距离误差低于2.5 m、7.5 m和0.6 m阈值（α）的像素所占百分比，定义为

$$\mathrm{PAG}_{\alpha}=\frac{\sum_{(i,j)\in \boldsymbol{G}\cap\hat{\boldsymbol{G}}} I\left(\left|h_{ij}-\hat{h}_{ij}\right|<\alpha\right)}{\sum_{(i,j)\in\hat{\boldsymbol{G}}} I\left((i,j)\in\hat{G}\right)} \tag{26}$$

Comp表示预测深度图中具有有效深度值像素所占的百分比，定义为

$$\mathrm{Comp}=\frac{\sum_{(i,j)\in G} I\left((i,j)\in D\right)}{\sum\left((i,j)\in(r,c)\right)} \tag{27}$$

式中，r和c分别代表图像像素的行数和列数。小于“3-interval”表示预测值与真实值之间的L_1误差小于3个深度间隔的像素所占百分比。由于LuoJia-MVS数据集空间分辨率为10 cm，推测小于“3-interval”相当于小于0.3 m。

推理时间表示从输入多视图遥感图像到完成深度图重建的整个过程所需的运行时间。

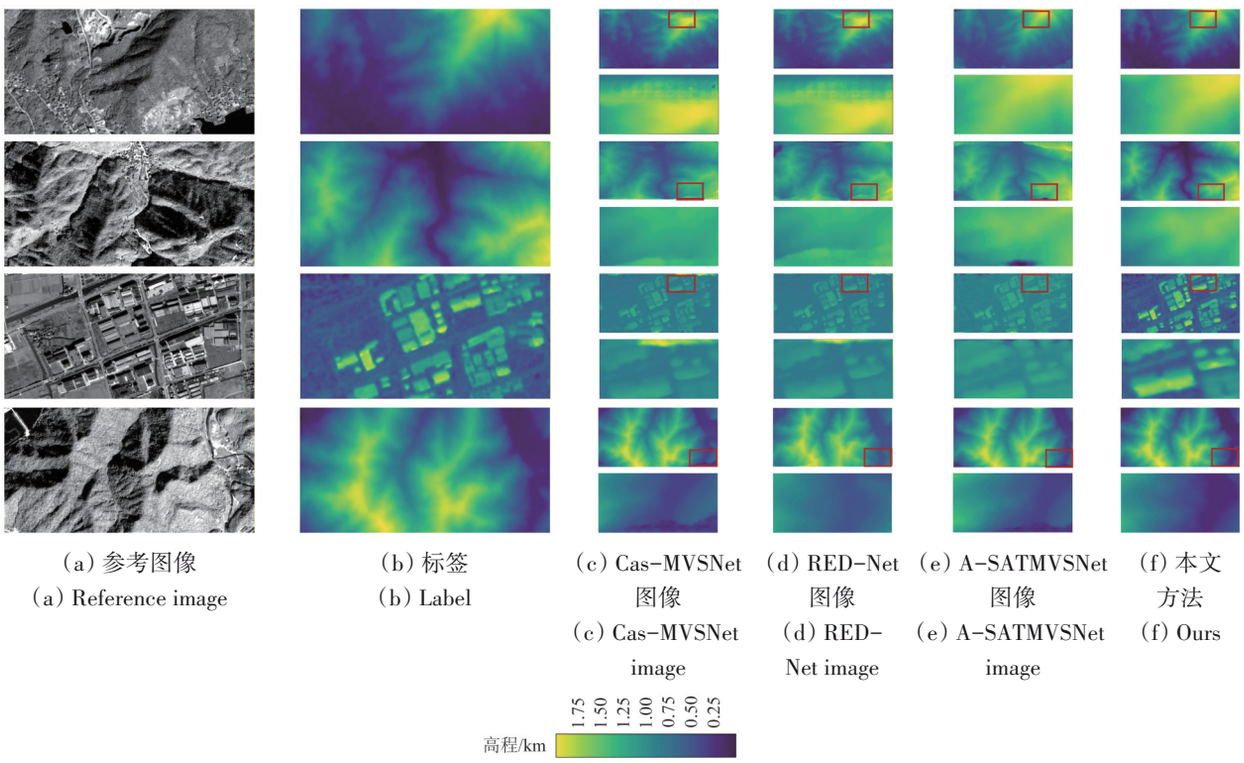
3.3 实验设置

本文采用PyTorch开源深度学习框架构建网络，在搭载Intel Core i7-9700 CPU、RTX 2080Ti显卡、16GB内存的计算机上进行实验。在WHU-TLC数据集训练阶段，将输入图像的分辨率设置为768×384，视图数$N=3$；LuoJia-MVS数据集上设置输入图像分辨率为784×368，图像的视图数N分别设置为3和5。各阶段平面扫描深度假设分别为48、32和8，深度间隔分别为4、2和1。网络训练时使用Adam优化器，优化参数设置为β_1=0.9和β_2=0.999，损失函数参数设置为λ_{DE}=0.5、λ_{ED}=0.5，各阶段损失权重分别为0.5，1.0和2.0。训练16个轮次，批量大小设置为1，初始学习率设置为0.001，在第10、12和14个轮次时，学习率减半。

3.4 数据集重建结果

将本文方法与MVSNet（Yao等，2018）、R-MVSNet（Yao等，2019）、Fast-MVSNet（Yu和Gao，2020）、Cas-MVSNet（Gu等，2020）、UCS-Net（Cheng等，2020）、RED-Net（Liu和Ji，2020）、A-SATMVSNet（Lin等，2023）、HDC-MVSNet（Li等，2023）、SA-SatMVS（Chen等，2024）方法在WHU-TLC与LuoJia-MVS数据集上进行实验对比，超参数设置与对比方法保持一致。

图7展示了Cas-MVSNet、RED-Net、A-SATMVSNet和本文提出方法共4种方法在WHU-TLC数据集上的深度预测结果对比。可见，其他方法处理复杂地形时会出现边缘细节丢失、纹理模糊和存在噪声等问题，而本文方法减少了重建深度图中的噪声干扰。在山地弱纹理和边缘区域重建效果要更加细腻，更接近真实值，在保持边缘结构的完整和精确度方面优于其他方法。



（a）参考图像 （a）Reference image　（b）标签 （b）Label　（c）Cas-MVSNet图像 （c）Cas-MVSNet image　（d）RED-Net图像 （d）RED-Net image　（e）A-SATMVSNet图像 （e）A-SATMVSNet image　（f）本文方法 （f）Ours

图7　不同模型在WHU-TLC数据集上的重建结果对比（每一组分上下2个小图，下面小图是上面图中红色矩形框的局部细节放大图）

Fig. 7　Comparative analysis of reconstruction performance across different models on the WHU-TLC dataset

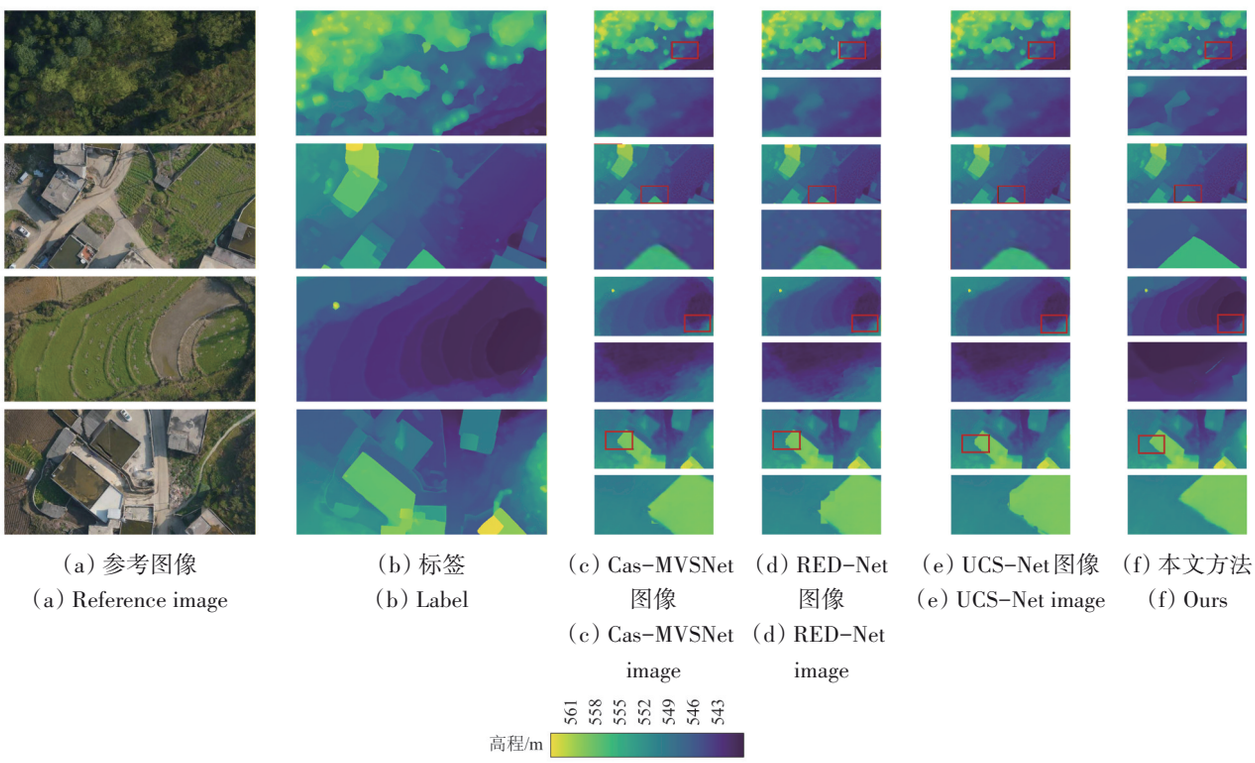
图8展示了4种方法在LuoJia-MVS数据集上深度预测可视化结果。这里选取丛林、房屋、梯田3种地形场景。从细节放大视图中可以看出，其他方法在边缘重建较为模糊和嘈杂，存在噪声干扰现象，而DiffusionMVS方法由于引入扩散约束模块，能够重建出更为清晰和平滑的边缘。显然，本文提出的方法在房屋建筑和林地区域的重建效果均优于其他对比方法，尤其是在房屋边缘区域，可有效抑制噪声并保持较高的重建精度，展现出良好的泛化能力和鲁棒性。

表1为本文方法与MVSNet、R-MVSNet、Fast-MVSNet、Cas-MVSNet、UCS-Net、RED-Net、A-SATMVSNet和SA-SatMVS在WHU-TLC数据集上的比较结果。可见，在MAE、RMSE、PA $G_{2.5\,m}$、PA $G_{7.5\,m}$和完整度评价指标上分别达到1.56、2.02、83.62%、96.76%和84.34%；比基准模型RED-Net分别提高28.11%、55.21%、9.49%、0.85%和2.52%；比A-SATMVSNet分别提高2.50%、0.98%、0.94%、0.28%和0.02%；比SA-SatMVS模型分别提高17.02%、46.70%、4.60%、0.14%和1.97%；本文方法在推理时间上相比A-SATMVSNet降低14.22%。总的来说，相较于部分方法，尽管DiffusionMVS在推理时间上略有增加，但在其他度量指标上实现整体提升。

表2给出在LuoJia-MVS数据集上本文方法与MVSNet、R-MVSNet、Fast-MVSNet、Cas-MVSNet、UCS-Net、RED-Net和HDC-MVSNet的比较结果。可以看出，将3个视角的图像作为输入时，本文方法在MAE、PA $G_{0.6\,m}$和<3-interval指标上分别达到0.086、98.8%和97.9%；比基准模型HDC-MVSNet分别提高3.37%、0.10%和0.10%。将5个视角的图像作为输入时，在MAE、PA $G_{0.6\,m}$和<3-interval指标上分别达到0.119、98.4%和97.4%；比基准模型HDC-MVSNet分别提高1.65%、0.10%和0.8%。从推理时间上看，尽管扩散约束模块运行效率较低，但本文方法在精度和效率之间取得更好的平衡，更适用于遥感影像立体重建任务。

图9展示的是本文方法与Cas-MVSNet、RED-Net在自建的油气站场数据集的深度预测结果对比。为评估本文方法在复杂工业场景中的重建性能，这里选取具有典型特征的输油管道、密集分布的油气储罐、栅格化围墙以及罐式集装箱车厢4个区域进行对比分析。实验结果表明，现有方法在几何细节重建方面存在几何细节丢失问题，如

管道连接处形变失真、密集储罐重建模糊等问题。本文方法通过自适应特征聚合模块来动态整合不同层次的特征实现更加精细化的几何特征重建，展现出更优的边缘保持能力，这验证了算法在未知场景下的泛化能力。



（a）参考图像 （a）Reference image　（b）标签 （b）Label　（c）Cas-MVSNet图像 （c）Cas-MVSNet image　（d）RED-Net图像 （d）RED-Net image　（e）UCS-Net图像 （e）UCS-Net image　（f）本文方法 （f）Ours

图8　不同模型在LuoJia-MVS数据集上的重建结果对比(每一组分上下两小图,下面小图是上面图中红色矩形框)

Fig. 8　Comparative analysis of reconstruction performance across different models on the LuoJia-MVS dataset

表1　不同方法在WHU-TLC数据集上定量对比

Table 1　Quantitative benchmarking of state-of-the-art methods on the WHU-TLC dataset

方法	MAE/m	RMSE/m	PA $G_{2.5\,m}$/%	PA $G_{7.5\,m}$/%	Comp/%	推理时间
MVSNet	2.30	4.87	63.50	93.80	80.33	15 min 02 s
R-MVSNet	2.23	4.67	63.90	95.00	81.68	13 min 57 s
Fast-MVSNet	2.24	4.58	74.27	95.87	82.02	3 min 52s
Cas-MVSNet	2.03	4.35	77.39	96.53	82.33	4 min 02s
UCS-Net	2.03	4.08	76.40	96.66	82.08	**3** min **47**s
RED-Net	2.17	4.51	74.13	95.91	81.82	9 min 15s
A-SATMVSNet	1.60	2.04	82.68	96.48	84.32	14 min 53s
SA-SatMVS	1.88	3.79	79.02	96.62	82.37	
本文方法	**1.56**	**2.02**	**83.62**	**96.76**	**84.34**	12 min 46s

注：数值加粗表示不同方法中的最高精度。

3.5　消融实验结果

为了验证本文的MFE-FPN模块、AFA模块和DCM模块对深度图重建结果的影响，在WHU-TLC数据集上分别对MFE-FPN模块、AFA模块、DCM模块和EDG模块进行消融实验，实验参数保持相同设置，结果见表3。具体结果如下：

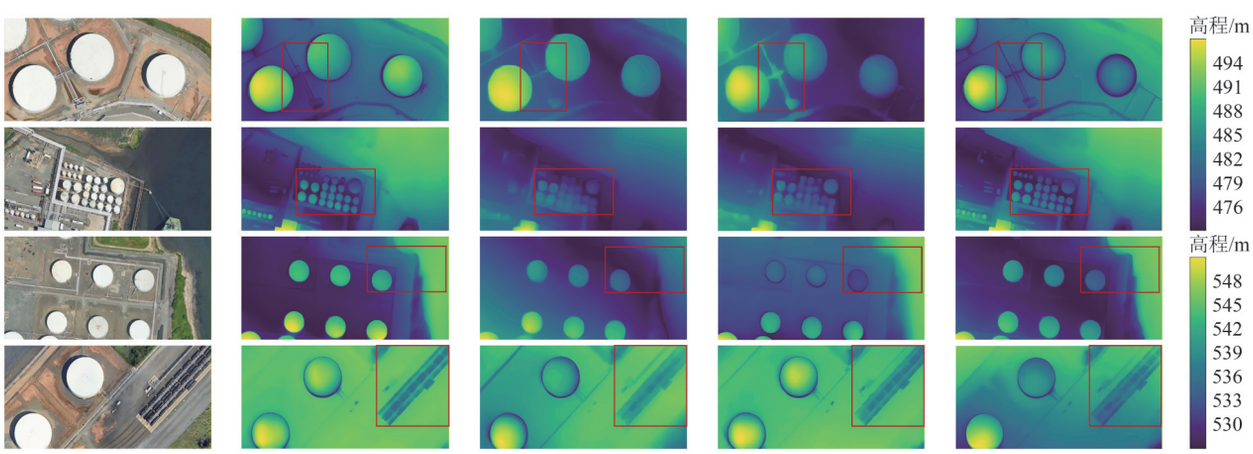
（1）为了验证MFE-FPN模块的有效性，去除MFE-FPN模块重新进行实验，实验结果如表3所示。可见，在MAE指标上N比N-MFE-FPN下降9.30%；RMSE指标上N比N-MFE-FPN下降38.41%；在PA $G_{2.5\,m}$和PA $G_{7.5\,m}$指标上分别提高6.95%和3.35%。如图10所示，引入MFE-FPN模块后，模型可提取到更加丰富的边缘特征，重建精度显著提高。结果表明MFE-FPN模块可提高模型对于多尺度遥感影像特征的表达能力。

表2 不同方法在LuoJia-MVS数据集上定量对比

Table 2 Quantitative benchmarking of state-of-the-art methods on the LuoJia-MVS dataset

输入视图数量	方法	MAE/m	PA $G_{0.6\,m}$/%	<3-interval/%	推理时间
三视图	MVSNet	0.172	96.1	92.4	2 min 26 s
	R-MVSNet	0.177	96.0	93.5	1 min 33 s
	Fast-MVSNet	0.194	95.7	92.0	**0** min **37** s
	Cas-MVSNet	0.103	98.4	97.1	1 min 06 s
	UCS-Net	0.113	98.3	97.3	0 min 57 s
	RED-Net	0.109	98.2	96.9	1 min 32 s
	HDC-MVSNet	0.089	98.7	97.8	
	本文方法	**0.086**	**98.8**	**97.9**	2 min 10 s
五视图	MVSNet	0.270	91.2	81.8	3 min 24 s
	R-MVSNet	0.259	92.3	86.7	2 min 21 s
	Fast-MVSNet	0.357	84.6	74.9	**0** min **56** s
	Cas-MVSNet	0.141	97.9	95.4	2 min 04 s
	UCS-Net	0.139	97.7	95.3	1 min 27 s
	RED-Net	0.156	94.9	90.5	2 min 46 s
	HDC-MVSNet	0.121	98.3	96.6	
	本文方法	**0.119**	**98.4**	**97.4**	3 min 16 s

注：数值加粗表示不同方法中的最高精度。



（a）参考图像 （b）标签 （c）Cas-MVSNet图像 （d）RED-Net图像 （e）本文方法

（a）Reference image （b）Label （c）Cas-MVSNet image （d）RED-Net image （e）Ours

不同方法重建结果差异明显的区域

图9 不同模型在自建数据集上的重建结果对比

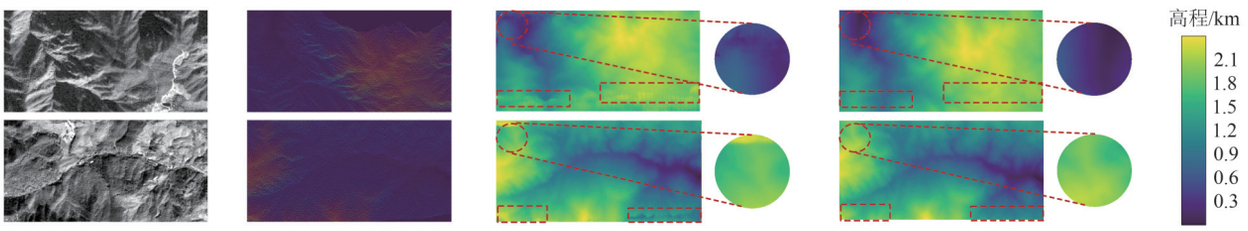
Fig. 9 Comparative analysis of reconstruction performance across different models on the self-built dataset

表3 消融实验结果

Table 3 Ablation study results

模型	MAE/m	RMSE/m	PA $G_{2.5\,m}$/%	PA $G_{7.5\,m}$/%
N-MFE-FPN	1.72	3.28	76.67	93.41
N-AFA	1.86	3.63	80.32	95.42
N-DCM	1.61	2.06	82.41	96.37
N-EDG	1.65	2.12	81.97	96.26
N	**1.56**	**2.02**	**83.62**	**96.76**

注：N-MFE-FPN、N-AFA、N-DCM和N-EDG分别表示不包含MFE-FPN模块、AFA模块、DCM模块和EDG模块的网络模型，*N*表示引入上述模块的网络模型。



（a）参考图像 （b）标签 （c）N-MFE-FPN图像 （d）*N*图像
（a）Reference image （b）Label （c）N-MFE-FPN image （d）*N* image

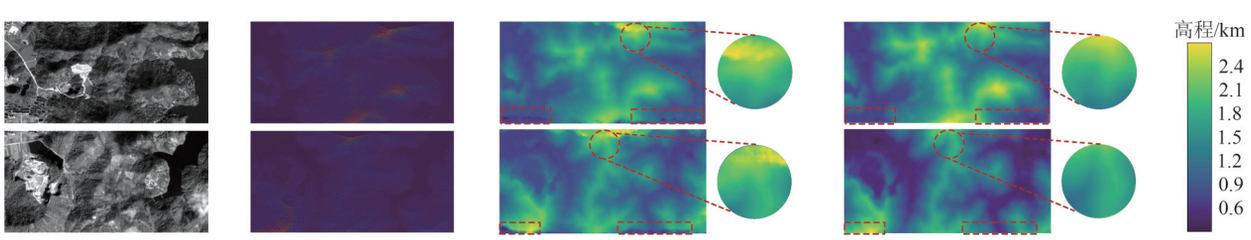
图10 MFE-FPN模块对深度图重建局部细节影响（图上方虚线圆圈以及下方左右2侧的虚线矩形框的区域是不同模型重建结果差异明显的区域）

Fig. 10 Impact of the MFE-FPN module on local detail reconstruction in depth maps

（2）为了验证AFA模块的有效性，去除AFA模块重新进行实验，实验结果如表3所示。可见，在MAE指标上*N*比N-AFA下降16.13%；RMSE指标上*N*比N-AFA下降44.35%；在PA $G_{2.5\,m}$和PA $G_{7.5\,m}$指标上分别提高3.30%和1.34%。如图11所示，引入AFA模块后，模型在边缘区域重建更为清晰，进一步提升模型的整体性能。

（3）为了验证DCM模块的有效性，去除DCM模块重新进行实验。其中N-DCM表示不包含DCM模块的网络模型，*N*表示引入DCM模块的网络模型。可以看出，在MAE指标上*N*比N-DCM下降3.11%；RMSE指标上*N*比N-DCM下降1.94%；在PA $G_{2.5\,m}$和PA $G_{7.5\,m}$指标上分别提高1.21%和0.39%。如图12所示，引入DCM模块后，有效减少了深度图中存在的噪声干扰，模型边缘重建精度显著提高，证明该模块设计的有效性。

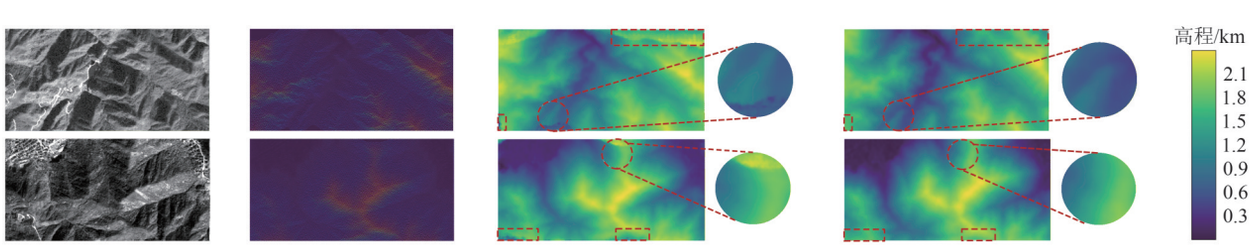
（4）为了验证EDG模块的有效性，去除EDG模块重新进行实验。可以看出，在MAE指标上*N*比N-EDG下降5.45%；RMSE指标上*N*比N-EDG下降4.72%；在PA $G_{2.5\,m}$和PA $G_{7.5\,m}$指标上分别提高1.65%和0.50%。



（a）参考图像 （b）标签 （c）N-AFA图像 （d）*N*图像
（a）Reference image （b）Label （c）N-AFA image （d）*N* image

图11 AFA模块对深度图重建局部细节影响（图上方虚线圆圈以及下方左右2侧的虚线矩形框的区域是不同模型重建结果差异明显的区域）

Fig. 11 Impact of the adaptive feature aggregation module on local detail reconstruction in depth maps



（a）参考图像 （b）标签 （c）N-DCM图像 （d）*N*图像
（a）Reference image （b）Label （c）N-DCM image （d）*N* image

图12 DCM模块对深度图重建局部细节影响（图上方虚线圆圈以及下方左右2侧的虚线矩形框的区域是不同模型重建结果差异明显的区域）

Fig. 12 Impact of the diffusion constrained module on local detail reconstruction in depth maps

4 结 论

为减少遥感影像立体重建中存在的噪声干扰，本文提出一种基于扩散约束的遥感影像立体重建算法。该方法融合了多尺度特征增强模块、自适应特征聚合模块和基于扩散约束的代价体优化模块，并在WHU-TLC和LuoJia-MVS公开数据集以及自建数据集上与多个基准方法（如Cas-MVSNet、RED-Net、UCS-Net等）进行了对比实验验证。主要结论如下：（1）MFE-FPN模块在特征金字塔网络基础上引入特征增强机制，有效提升了网络对遥感影像多尺度特征的提取与表征能力。（2）AFA模块通过动态整合不同层次的特征，可增强对目标边缘等深度细节的捕获能力，改善了边缘区域的深度估计精度。（3）扩散约束模块通过扩散机制优化深度值分布以抑制噪声，并结合边缘引导的Transformer可提升深度图的边缘重建质量。

本文方法能够解决遥感影像深度重建中存在的特征匹配精度低、预测深度图存在噪声干扰和边缘重建不完整等问题。未来考虑将分割一切模型SAM（Segment Anything Model）大模型技术（Kirillov等，2023）嵌入MVS网络中，通过SAM分割后的语义信息优化MVS的匹配过程，进一步提高模型重建效率和精度。

参考文献（References）

Barnes C, Shechtman E, Finkelstein A and Goldman D B. 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics, 28(3): 24 [DOI: 10.1145/1531326.1531330]

Chen X L, Diao W H, Zhang S, Wei Z W and Liu C B. 2024. SA-SatMVS: slope feature-aware and across-scale information integration for large-scale earth terrain multi-view stereo. Remote Sensing, 16(18): 3474 [DOI: 10.3390/rs16183474]

Cheng S, Xu Z X, Zhu S L, Li Z W, Li L E, Ramamoorthi R and Su H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 2521-2531 [DOI: 10.1109/CVPR42600.2020.00260]

Dong R M, Yuan S, Luo B, Chen M X, Zhang J X, Zhang L X, Li W J, Zheng J P and Fu H H. 2024. Building bridges across spatial and temporal resolutions: reference-based super-resolution via change priors and conditional diffusion model//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 27674-27684 [DOI: 10.1109/CVPR52733.2024.02614]

Gallup D, Frahm J M, Mordohai P, Yang Q X and Pollefeys M. 2007. Real-time plane-sweeping stereo with multiple sweeping directions//2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis: IEEE: 1-8 [DOI: 10.1109/CVPR.2007.383245]

Gao J, Liu J and Ji S P. 2021. Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 6128-6137 [DOI: 10.1109/ICCV48922.2021.00609]

Gao J, Liu J and Ji S P. 2023. A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images. ISPRS Journal of Photogrammetry and Remote Sensing, 195: 446-461 [DOI: 10.1016/j.isprsjprs.2022.12.012]

Gu X D, Fan Z W, Zhu S Y, Dai Z Z, Tan F T and Tan P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 2492-2501 [DOI: 10.1109/CVPR42600.2020.00257]

Han L T, Zhao Y C, Lv H Y, Zhang Y S, Liu H L and Bi G L. 2022. Remote sensing image denoising based on deep and shallow feature fusion and attention mechanism. Remote Sensing, 14(5): 1243 [DOI: 10.3390/rs14051243]

Heo S and Lee S. 2024. Denoising diffusion for multi-view stereo//2024 International Conference on Electronics, Information, and Communication (ICEIC). Taipei, China: IEEE: 1-3 [DOI: 10.1109/ICEIC61013.2024.10457167]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc.: 6840-6851

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Ibrahimli N, Ledoux H, Kooij J F P and Nan L L. 2023. DDL-MVS: depth discontinuity learning for multi-view stereo networks. Remote Sensing, 15(12): 2970 [DOI: 10.3390/rs15122970]

Khan N, Kim M H and Tompkin J. 2021. Differentiable diffusion for dense depth estimation from multi-view images//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE: 8908-8917 [DOI: 10.1109/CVPR46437.2021.00880]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Li J Y, Huang X, Feng Y J, Ji Z, Zhang S L and Wen D W. 2023. A hierarchical deformable deep neural network and an aerial image benchmark dataset for surface multiview stereo reconstruction. IEEE Transactions on Geoscience and Remote Sensing, 61: 5600812 [DOI: 10.1109/TGRS.2023.3234694]

Li Z Y, Li Z Q, Cui Z P, Pollefeys M and Oswald M R. 2024. Sat2Scene: 3D urban scene generation from satellite images with diffusion//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 7141-7150 [DOI: 10.1109/CVPR52733.2024.00682]

Lin L, Zhang Y B, Wang Z J, Zhang L L, Liu X F and Wang Q Q. 2023. A-SATMVSNet: an attention-aware multi-view stereo matching network based on satellite imagery. Frontiers in Earth Science, 11: 1108403 [DOI: 10.3389/feart.2023.1108403]

Liu J, Gao J, Ji S P, Zeng C, Zhang S Y and Gong J Y. 2023a. Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 204: 42-60 [DOI: 10.1016/j.isprsjprs.2023.08.015]

Liu J and Ji S P. 2020. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 6049-6058 [DOI: 10.1109/CVPR42600.2020.00609]

Liu N N, Wang P H, Xiang S Y, Gu N N and Wang F. 2023b. RS-MVSNet: inferring the earth's digital surface model from multi-view optical remote sensing images//IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society. Singapore: IEEE: 1-7 [DOI: 10.1109/IECON51785.2023.10311909]

Luo H T, Zhang J M, Liu X F, Zhang L L and Liu J Y. 2024. Large-scale 3D reconstruction from multi-view imagery: a comprehensive review. Remote Sensing, 16(5): 773 [DOI: 10.3390/rs16050773]

Mao Y Q, Bi H B, Xu L Y, Chen K Q, Wang Z R, Sun X and Fu K. 2024. SDL-MVS: view space and depth deformable learning paradigm for multiview stereo reconstruction in remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62: 5641518 [DOI: 10.1109/TGRS.2024.3464574]

Merrell P, Akbarzadeh A, Wang L, Mordohai P, Frahm J M, Yang R G, Nister D and Pollefeys M. 2007. Real-time visibility-based fusion of depth maps//2007 IEEE 11th International Conference on Computer Vision. Rio de Janeiro: IEEE: 1-8 [DOI: 10.1109/ICCV.2007.4408984]

Schönberger J L and Frahm J M. 2016. Structure-from-motion revisited//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 4104-4113 [DOI: 10.1109/CVPR.2016.445]

Shao R Z, Zheng Z R, Zhang H W, Sun J X and Liu Y B. 2022. DiffuStereo: high quality human reconstruction via diffusion-based stereo using sparse cameras//17th European Conference on Computer Vision. Tel Aviv: Springe: 702-720 [DOI: 10.1007/978-3-031-19824-3_41]

Song J M, Meng C L and Ermon S. 2022. Denoising diffusion implicit models. arXiv preprint arXiv: 2010.02502 [DOI: 10.48550/arXiv.2010.02502]

Toker A, Eisenberger M, Cremers D and Leal-Taixé L. 2024. SatSynth: augmenting image-mask pairs through diffusion models for aerial semantic segmentation//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 27685-27695 [DOI: 10.1109/CVPR52733.2024.02615]

Wang L, Jia J L and Dai H L. 2024. OrientedDiffDet: diffusion model for oriented object detection in aerial images. Applied Sciences, 14(5): 2000 [DOI: 10.3390/app14052000]

Wei Z Z, Zhu Q T, Min C, Chen Y S and Wang G P. 2021. AA-RMVSNet: adaptive aggregation recurrent multi-view stereo network//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE: 6167-6176 [DOI: 10.1109/ICCV48922.2021.00613]

Wen Y H, Ma X P, Zhang X K and Pun M O. 2024. GCD-DDPM: a generative change detection model based on difference-feature guided DDPM. IEEE Transactions on Geoscience and Remote Sensing, 62: 5404416 [DOI: 10.1109/TGRS.2024.3381752]

Woo S, Park J, Lee J Y and Kweon, I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Wu B J and Huang H. 2020. Survey on 3D reconstruction of transparent objects. Journal of Computer-Aided Design and Computer Graphics, 32(2): 173-180 (吴博剑, 黄惠. 2020. 透明物体的三维重建综述. 计算机辅助设计与图形学学报, 32(2): 173-180) [DOI: 10.3724/SP.J.1089.2020.18101]

Wu Z T, Xiao M Q, Fang C and Lin Z C. 2024. Designing universally-approximating deep neural networks: a first-order optimization approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(9): 6231-6246 [DOI: 10.1109/TPAMI.2024.3380007]

Xu Q S and Tao W B. 2019. Multi-scale geometric consistency guided multi-view stereo//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 5478-5487 [DOI: 10.1109/CVPR.2019.00563]

Yan H B, Xu F Q, Huang L E, Liu C B and Lin C X. 2023. Review of multi-view stereo reconstruction methods based on deep learning. Optics and Precision Engineering, 31(16): 2444-2464 (鄢化彪, 徐方奇, 黄绿娥, 刘词波, 林初欣. 2023. 基于深度学习的多视图立体重建方法综述. 光学精密工程, 31(16): 2444-2464) [DOI: 10.37188/OPE.20233116.2444]

Yao Y, Luo Z X, Li S W, Fang T and Quan L. 2018. MVSNet: depth inference for unstructured multi-view stereo//15th European Conference on Computer Vision. Munich: Springer: 785-801 [DOI: 10.1007/978-3-030-01237-3_47]

Yao Y, Luo Z X, Li S W, Shen T W, Fang T and Quan L. 2019. Recurrent MVSNet for high-resolution multi-view stereo depth inference//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 5520-5529 [DOI: 10.1109/CVPR.2019.00567]

Yu D W, Ji S P, Liu J and Wei S Q. 2021. Automatic 3D building reconstruction from multi-view aerial images with deep learning. ISPRS Journal of Photogrammetry and Remote Sensing, 171: 155-170 [DOI: 10.1016/j.isprsjprs.2020.11.011]

Yu Z H and Gao S H. 2020. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 1946-1955

[DOI: 10.1109/CVPR42600.2020.00202]

Zhang S, Wei Z W, Xu W J, Zhang L L, Wang Y, Zhang J M and Liu J Y. 2024. Edge aware depth inference for large-scale aerial building multi-view stereo. ISPRS Journal of Photogrammetry and Remote Sensing, 207: 27-42 [DOI: 10.1016/j.isprsjprs.2023.11.020]

Zhang S, Wei Z W, Xu W J, Zhang L L, Wang Y, Zhou X and Liu J Y. 2023a. DSC-MVSNet: attention aware cost volume regularization based on depthwise separable convolution for multi-view stereo. Complex and Intelligent Systems, 9(6): 6953-6969 [DOI: 10.1007/s40747-023-01106-3]

Zhang S, Xu W J, Wei Z W, Zhang L L, Wang Y and Liu J Y. 2023b. ARAI-MVSNet: a multi-view stereo depth estimation network with adaptive depth range and depth interval. Pattern Recognition, 144: 109885 [DOI: 10.1016/j.patcog.2023.109885]

Zhou L Y, Zhang Z, Jiang H Q, Sun H, Bao H J and Zhang G F. 2021. DP-MVS: detail preserving multi-view surface reconstruction of large-scale scenes. Remote Sensing, 13(22): 4569 [DOI: 10.3390/rs13224569]

DiffusionMVS：Multi−view stereo reconstruction algorithm for remote sensing image based on diffusion constraints

LIAN Yuanfeng[1,2], WANG Sen[1]

1. College of Artificial Intelligence, China University of Petroleum, Beijing 102249, China;
2. Beijing Key Laboratory of Petroleum Data Mining, Beijing 102249, China

Abstract：Large-scale 3D scene reconstruction based on remote sensing images provides critical support for smart city development, map navigation, virtual reality, and digital twin systems. Existing 3D reconstruction algorithms predominantly rely on feature matching techniques and demonstrate satisfactory performance in small-scale or structurally simple scenes. Given the intricate terrain features and noise interference in complex or large-scale environments, significant challenges, such as suboptimal reconstruction accuracy and incomplete modeling, exist. These challenges hinder the effectiveness of these methods. Therefore, this study proposes a diffusion-constrained multiview stereo network comprising a multiscale feature enhancement feature pyramid network (MFE-FPN), an adaptive feature aggregation module (AFA), and a diffusion-constrained module (DCM) to address the issues of low-feature matching accuracy, high noise in predicted depth maps, and incomplete edge reconstruction in multiview stereo for remote sensing images.

The proposed method consists of several steps. First, the network takes N multiview remote sensing images as input, with the first image serving as the reference and the remaining N-1 as source images. It adopts a three-stage coarse-to-fine strategy to predict depth maps progressively. The network utilizes the MFE-FPN module to extract multiscale features from the input images, thereby generating hierarchical feature representations. Second, the top-level features from the FPN are mapped through an edge-aware network to compute edge-aware features, which are subsequently fused with the multiscale features. Third, an AFA is designed to aggregate the multiscale features, thereby forming a matching cost volume. Fourth, a diffusion constraint module is introduced to integrate cost volume features with edge-aware features. Fifth, an edge-guided transformer is employed to enhance the representation of edge details during the denoising stage. Sixth, the cost volume features are regularized and regressed to estimate depth, resulting in the final reconstructed depth map. Seventh, an edge-aware loss function is constructed during training to preserve the edge information in the predicted depth maps effectively.

Experimental results show that compared with other methods, the DiffusionMVS network shows an improved mean absolute error metric on the WHU-TLC and LuoJia-MVS datasets by 28.11% and 3.37%, respectively, thereby demonstrating superior reconstruction performance. However, in terms of inference time, the proposed method does not achieve the best performance because of the relatively low operational efficiency of the diffusion constraint module. Nevertheless, it achieves an optimal balance between accuracy and efficiency, thereby making it highly suitable for remote sensing stereo reconstruction tasks. The results on the self-constructed dataset of oil and gas stations verify the model's capability to reconstruct detailed geometric features. This capability benefits from the model's excellent performance in edge preservation and generalization in unseen scenarios. Moreover, ablation experiment results confirm that the proposed MFE-FPN, AFA, and DCM modules can effectively enhance the accuracy of depth map reconstruction.

The proposed diffusion-constrained multiview stereo network significantly improves edge-processing capability and overall reconstruction accuracy through a multiscale feature enhancement module and a diffusion constraint module. Results indicate the model is well-suited for reconstructing mountains, forests, and buildings, because of its superior performance on weak-texture regions and depth map denoising challenges. It effectively addresses the reduced reconstruction accuracy of remote sensing images under noise interference. Future work will explore incorporating the Segment Anything Model into the MVS framework to leverage its rich semantic information, thereby refining the matching process and further improving reconstruction efficiency and accuracy.

Key words：remote sensing images, multi-view stereo, multi-scale feature extraction, adaptive feature aggregation, diffusion model, edge-guided transformer

Supported by National Natural Science Foundation of China (No. 61972353); China National Petroleum Corporation-China University of Petroleum (Beijing) Strategic Cooperation Science and Technology Project (No. ZLZX2020-05)