基于生成模型的无监督多视点立体视觉网络

潘宇轩[1]，金锐[1]，刘雨[1]，张琳[1,2]

(1. 北京邮电大学人工智能学院，北京 100876；
2. 北京市大数据中心，北京 100086)

摘　　要：现有的多视点立体视觉研究利用深度估计算法，通过建立物理世界与数字世界的映射关系来实现立体表征。基于有监督学习的神经网络算法通过训练能够取得准确且高保真的三维重建结果。然而，由于缺乏深度先验信息且图像具备大视场的特性，面向自然场景的视觉重建仍然具有挑战性。研究应用无监督学习网络和基于语义优化的神经辐射场(NeRF)渲染，在没有先验信息的情况下实现对自然采集的多视点图像的深度估计。首先通过无监督学习无参考地生成多视点图像初步的深度信息，进一步在独立的 NeRF 模型中，利用扩散模型建立表面语义渲染损失来实现细粒度的三维表征。在基准数据集上的实验结果表明，该方法与其他最先进的方案相比整体重建的指标平均提高了 24.6%；在宽基线数据集的泛化性能验证中，该方法将现有方法测得的重建误差最多降低了 40.8%。

关　键　词：无监督深度学习；多视点立体视觉；三维重建；神经辐射场；深度优化

中图分类号：TP 391.41　　**DOI**：10.11996/JG.j.2095-302X.2026010029

文献标识码：A　　**文 章 编 号**：2095-302X(2026)01-0029-10

Generative model based unsupervised multi-view stereo network

PAN Yuxuan[1], JIN Rui[1], LIU Yu[1], ZHANG Lin[1,2]

(1. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Beijing Big Data Center, Beijing 100086, China)

Abstract: Existing research on multi-view stereo scheme utilizes depth-estimation algorithms to achieve stereo representation by establishing a mapping relationship between the physical and digital worlds. Supervised learning-based neural networks have achieved accurate and high-fidelity 3D reconstruction results through training. However, in-the-wild visual reconstruction remains challenging due to the lack of rendered depth priors and wide-baseline characteristics of images. A novel system was proposed to obtain optimized depth for naturally collected multi-view images without prior information by applying an unsupervised learning network and semantically optimized Neural Radiation Field (NeRF) rendering. First, preliminary depth information for wild multi-view images were produced without ground truth based on unsupervised deep learning. Subsequently, in a separate NeRF module, a diffusion model was used to construct a surface semantic rendering loss, enabling a fine-grained volumetric representation. Experimental results on the benchmark dataset validated the performance of the proposed system by improving an average of 24.6% of the overall metrics, compared with other state-of-the-art schemes. A novel wild wide-baseline dataset was also applied to verify the generalization performance, and the proposed system reduced the reconstruction error by up to 40.8% compared with all methods.

Keywords: unsupervised deep learning; multi-view stereo; 3D reconstruction; neural radiation field; depth optimization

收稿日期：2025-04-29；**定稿日期**：2025-06-28；**通信作者**：张琳，E-mail：zhanglin@bupt.edu.cn

Received：29 April，2025；**Finalized**：28 June，2025；**Corresponding author**：ZHANG Lin，E-mail：zhanglin@bupt.edu.cn

基金项目：国家重点研发计划(2023YFB2704500)；北京市自然科学基金(4222033)

Foundation items：National Key Research and Development Program of China (2023YFB2704500); Beijing Natural Science Foundation (4222033)

在三维计算机视觉技术中，多视点立体(Multi-View Stereo, MVS)系统旨在从多视点图像和经过校准的相机参数中计算出深度图，从而恢复真实场景的三维模型。随着深度学习在计算机视觉领域中表现出卓越的性能，将相关神经网络模型应用到 MVS 系统中，已在精确三维重建方面取得了显著进展[1]。

已有研究开发了基于 RGB 传感器的有监督深度学习方法，从原始 RGB 数据中创建立体模型。MVSNet[2]及其后续工作[3]引入了矢量化匹配损失，从而在卷积神经网络(Convolutional Neural Networks, CNN)中引入了四维损失函数实现更精准的立体匹配。这些有监督方法比较依赖具有真实深度先验的训练数据集，但先验数据在面向自然场景的三维重建中很难获取。为了在自然场景中进行稳健且完整的表面重建，研究人员转向无监督学习的研究，以实现在不需要任何深度先验的情况下提供准确的三维重建结果。DS-MVSNet[4]和 CL-MVSNet[5]提出了基于无监督学习的 MVS 网络，在不依赖真实三维训练数据的情况下从输入的多视点图像中推导深度图。然而，无监督方法中的光度一致性假设可能会被相机姿态变化、物体遮挡或自然图像的反射特征等因素破坏，造成测试数据和自然数据之间的性能差距，所以模型不能很好地迁移到任意自然数据的三维重建当中。

为了解决理论模型与实际应用之间的差异，具有隐式表示的神经渲染是三维重建方面有力的技术补充。神经辐射场(Neural Radiance Fields, NeRF)[6]以具有隐式表示的神经渲染模型出发，使用神经网络对三维空间光线进行学习，通过光线追踪对整个空间进行全局感知，在重建新视点方面有出色的表现。尽管 NeRF 以完全无监督的方式从体密度(Volume density)中获得空间的深度，但对于三维重建的高精度需求而言获得的深度还不够准确。如王道累等[7]通过将 NeRF 技术集成到 MVS 系统中来缓解这个问题。针对一个场景使用大量已知相机参数的图像进行隐式建模，利用源视点信息从中心的参考视点角度重建出清晰的场景，增强三维重建的泛化能力。

上述基于参考/源视点的 NeRF 监督的 MVS 方案，由于自然场景的大视场宽基线特性，其性能仍具备可提升的空间。在大规模重建中，单个视点无法完全覆盖所收集的场景，神经网络无法生成参考视点之外的内容。因此，需要在后期处理中将外部内容合成到参考视点中，和 NeRF 擅长的视点合成功能相匹配。研究探索将无监督 MVS 网络和 NeRF 设计为独立模块，针对任意自然数据以实现更完整的重建结果。所提出的系统利用了两者在三维重建中的优势，其中 MVS 网络估计主要对象的精确深度，而 NeRF 对其进行细化，以高精度补充周边对象的深度。考虑到边缘嵌入式人工智能技术能够满足三维重建服务的低延迟和高吞吐量需求[8]，所提出系统可以分别在云端和边缘应用独立的 MVS 和 NeRF，有利于减轻计算负担并减少推理时间。

本文提出了一种基于无监督深度学习的三维重建系统，以模块化结构实现自然图像的精确深度估计和立体匹配，主要贡献和创新为：

1) 提出了一种三维重建中处理多视点图像的无监督深度学习 MVS 网络，考虑了与视点相关的光度效应，从数据中生成深度图，然后通过立体重建生成体表示；

2) 提出了一个独立的基于 NeRF 的深度优化模块，引入关键点的深度监督损失来细化 MVS 网络估计的深度图，在竞争基准上的实验结果证明了本系统的有效性；

3) 引入了一个基于扩散模型的语义学习损失，实现视觉一致性下的三维场景感知，优化神经网络中的隐式函数，准确表达空间中的体积密度和颜色观测。

1 相关工作

在计算机视觉中，实现稳健三维重建是一项基础任务。现有的研究中包含了数学建模和基于深度学习网络等建立的三维模型的方法。

1.1 传统重建方法

传统立体匹配方法的流程聚焦于三维点，通常从一组稀疏的匹配关键点开始，并采用传播策略逐步使重建结果变得密集。刘鑫等[9]依靠 RGB(D)图像序列来重建几何模型。然而，基于关键点的工作受到并行数据处理能力的限制。研究人员通过从这些点计算深度图，然后将深度图进一步融合成点云可解决这个问题。SCHÖNBERGER 等[10]提出了 Colmap，使用手工特征并联合估计逐像素的视点选择、深度图和表面法线，以利用光度和几何先验信息。

1.2 基于深度学习的重建方法

最近，深度学习在立体视觉中表现出卓越的性能。继文献[2]的代表性工作之后，M3VSNet[3]提出了一种新颖的无监督多指标网络，用于在无任何监督的情况下进行密集点云重建。ShadowPatch[11]将光度立体视觉与立体深度估计相结合，使深度估计

更加稳健。LIANG 等[12]在此基础上进一步引入语义分割模型，根据分割的内容和深度共享边界的特点，优化立体估计的连续性。GST-MVS[13]利用图像纹理和对象深度之间的内在关系来增强深度估计，优化了 3D 重建过程。RobustMVS[14]引入了深度聚类引导的白化损失，在保持不同视图之间的特征一致性下从视点特定的风格信息中去关联多视图特征。一些工作从高质量数据库中学习数据驱动的先验知识，并从宽基线多视点相机中重建三维模型。SSC-MVS[15]建立了伪序列深度实现无监督学习，利用 2 个不同增强输入之间的深度一致性实现鲁棒的重建结果。DI-MVSNet[16]提出从粗到细的学习框架，使用深度感知迭代器将上下文引导的深度几何有效地整合到损失函数中。Unsup-MVS[17]提出了一种基于端到端学习的无监督框架：将源视点的图像根据其预测的深度图反向投影到参考视点。RENDLE 等[18]提出了一种系统设计，能够以更高的帧率进行体素化头像重建，克服了商用 RGB-D 相机的局限性。CT-MVSNet[19]将远程上下文聚合和全局特征交互进行扩展，可在不同分辨率下实现高效的特征匹配，对不准确的相机参数更具鲁棒性。

1.3　基于 NeRF 的重建方法

使用 NeRF[6]的深度图估计方法激发了许多 MVS 的优化研究。RC-MVSNet[20]提出了一种带有神经渲染的新颖方法，以解决视点间对应关系的模糊性问题，施加深度渲染一致性损失来约束几何特征，以减轻遮挡问题。DENG 等[21]提出了深度监督的 DS-NeRF，通过使用相机参数和 SfM (Structure from Motion)获得的稀疏三维点云来训练 NeRF，从而从少量图像中生成任意视角的图像。TOSI 等[22]探索了将 NeRF 用作稳健立体系统的新监督源，无需任何真实标签即可轻松训练 MVS 网络。MVCPS-NeuS[23]引入多视图约束光度立体理论，将表面法线估计的模糊度以 NeRF 全局线性计算，估计详细的三维模型。ZHU 等[24]引入了基于注意力机制和神经体渲染的 MVS 网络，使网络能够学习超出代价体表示的场景几何信息。NerfingMVS++[25]通过微调 Colmap 重建，采用自适应的深度先验来监控体数据绘制的采样过程，对渲染图像进行误差计算获得的每像素置信度图进一步提高深度质量。ITO 等[26]将 Colmap 和 NeRF 相结合来估计物体的深度图，将 NeRF 用作独立的深度图细化模块。ZHU 和 CHEN[27]在输入视点之间建立深度感知一致性，以加强视点之间的交互并缓解重建网络的过拟合问题。

受精确的神经渲染框架的启发，本文对基于无监督学习的 MVS 流程进行了改进，以实现稳健且通用的重建。通过使用基于 NeRF 的方法优化网络，利用深度和颜色渲染一致性损失来约束物体边界附近的几何特征，得到高精度的深度图以及相应的三维模型。

2　本文方法

本系统的整体架构遵循基于深度学习的立体视觉架构，如图 1 所示。该系统主要包含 2 部分，模块(1)为无监督多视点立体网络，由于现实世界场景无法提供准确的深度先验，需首先利用图像之间的对应关系实现相机参数的归一化表示，进一步将数据输入到无监督学习 MVS 网络中实现初步的深度估计；模块(2)为基于扩散模型优化的 NeRF 网络，利用扩散模型感知空间信息，推动 NeRF 实现更精确完整的深度估计，基于生成式深度学习最终提升三维重建的精度。

2.1　无监督多视点立体网络

给定总共 N 个视点作为网络的输入，将其分为 1 个参考视点和 $N-1$ 个源视点。考虑到系统的整体性，首先使用 SfM 方法[28]计算各视点的相机参数。通过将参考视点设置为世界坐标系，源视点可以通过空间关系进行数学建模，以充分利用视点数据之间的角度域或空间域相关性。

每个视点输入到对应的神经网络编码器中，各编码器共享相同的结构和权重，从而提取所有视点图像的二维特征，并通过可微单应性(Differentiable homography)理论[12]将源视点的特征扭曲(Warp)映射到参考视点，从而构建三维损失函数。然后对所得到的特征映射进行正则化，以获得整体的学习损失，并利用全连接概率生成深度图的预测结果。基于投影的坐标变换，可以对所有视点预测得到的深度图进行滤波和融合，以获得全面的深度估计结果。

具体来说，本文提出的无监督学习 MVS 网络基于深度卷积神经网络构建，通过多层卷积逐步细化的方式预测深度图。在本网络中，首先获取集合 $V=\{v_n|n=1,2,\cdots,N\}$ 中的所有 N 个视点 v_n 作为输入。无监督深度学习网络强制参考视点与其他源视点之间保持光度一致性，其关键思想是在计算损失函数时，最大化参考视点 v_R 与任何源视点 $v_s \subseteq v_n$ 扭曲到参考视点后的相似度。因此，网络需要分别依赖输入视点对应相机的内参 K 和外参$[R,t_v]$，对集合$\{v_n\}$进行编码，该网络生成逐像素特征并构建一个扭曲特征体$\{\boldsymbol{F}_j \mid j=1,2,\cdots,N\}$。

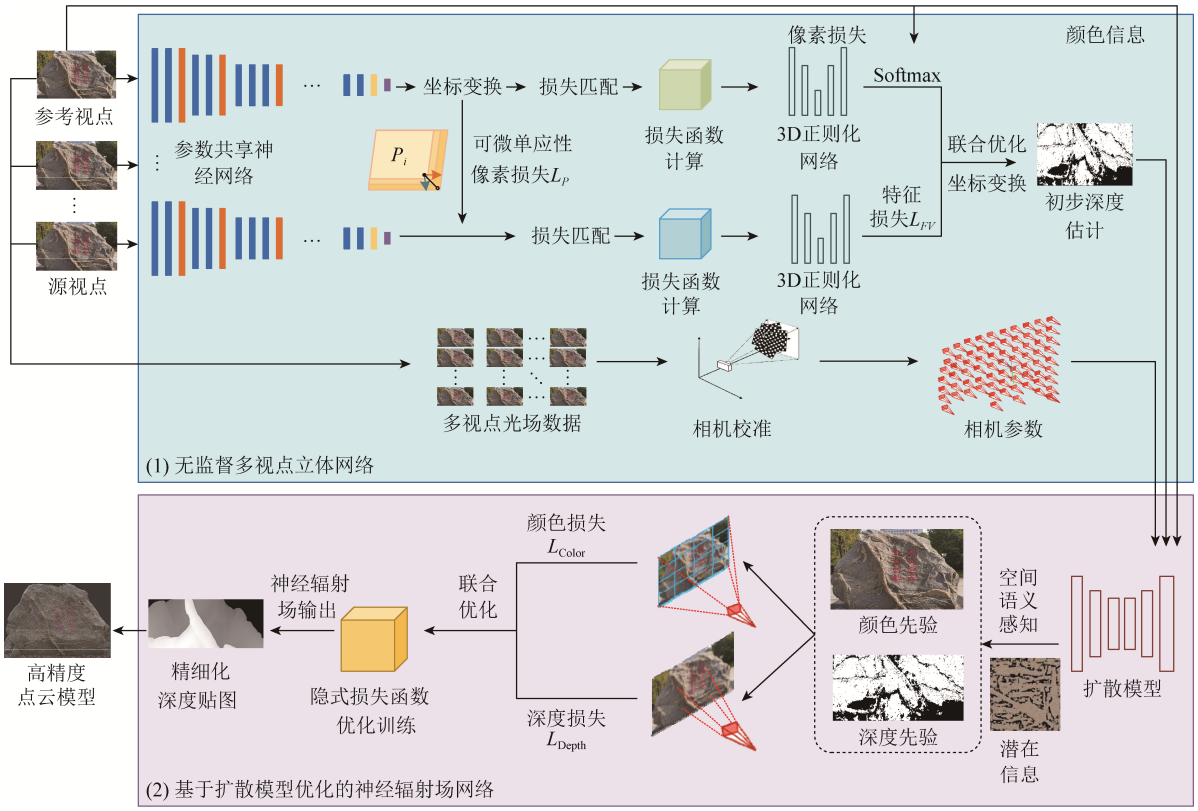

图 1 系统架构

Fig. 1 System architecture

在齐次坐标表示中，参考视点中的任何像素 p_i 都可以通过反向扭曲变换到源视点，即

$$\hat{p}_i = K \cdot T\left(\boldsymbol{D}_i K^{-1} p_i\right) \tag{1}$$

式中：T 表示从参考视点到源视点的相对变换；$\boldsymbol{D}_i$ 表示参考视点中 p_i 的预测深度值。扭曲后的像素 $\hat{p}_i$ 可通过在相应位置对其他源视点的颜色值进行双线性采样来形成一个新视点 $\hat{v}_s^j$ 中的像素，即满足 $\hat{v}_s^j(p_i) = v_j(\hat{p}_i)$。实际上，可以通过特征体 $\boldsymbol{F}_j$ 计算 $\hat{p}_i$，进而反过来求解深度值 $\boldsymbol{D}_i$。参考视点 v_R 还可以提取更多语义层面信息，以构建基于特征的损失函数。依上述定义，扭曲像素 $\hat{p}_i$ 与源视点中的特征 $\bar{F}_s^j$ 相关。为了减少随机误差，通过计算特征方差，将来自 v_R 的所有扭曲特征体 $\boldsymbol{F}_j$ 在视点间融合成一个公共特征体，即

$$\boldsymbol{F}_c = \frac{1}{m_j}\sum_{j=1}^{N}\left(F_R^j - \bar{F}_S^j\right)^2 \tag{2}$$

在 MVS 网络中，输出估计的深度之前需应用一个深度滤波过程，以过滤冗余的像素。使用二值掩码 M_j 对深度点施加更多约束，屏蔽投影到源视点边界之外的无效像素，有助于降低网络学习的复杂度，细化深度估计结果。应用掩码过程中，对应的二范数及梯度通过特殊的方式进行计算[5]，即

$$L_2(e) = \left(\sum_{i=1}^{n} e_i^{1/2}\right)^2,\quad \frac{\partial L_2(e)}{\partial e_i} = \left(\sum_{i=1}^{n} e_i^{1/2}\right)\cdot e_i^{-1/2} \tag{3}$$

式中：e 表示在光度一致性损失计算时对于网络生成图像和原始图像在特定像素处的距离，表征该范式具备逐像素计算的特点。

以最小化损失函数的值为训练目标，将多视点数据集输入到无监督网络进行迭代，实现深度图生成。基于视点间的相关关系，损失函数包括了二维平面点的损失和超越像素层面的特征对应匹配不同视点的损失，即

$$\begin{cases} \arg\min L = \lambda_1 L_{\text{photo}} + \lambda_2 L_{\text{SSIM}} + \lambda_3 L_{\text{Smooth}} + \lambda_4 L_{FV} \\ \text{s.t. } L_{\text{photo}} = \sum_{j=1}^{N}\dfrac{1}{m_j}\left\|\beta_1\left(\hat{v}_s^j - v_R\right) + \left(\beta_2\left(\nabla\hat{v}_s^j - \nabla v_R\right)\right)M_j\right\|_2 \\ L_{\text{SSIM}} = \sum_{j=1}^{N}\dfrac{1}{m_j}\left\|\dfrac{1-SSIM\left(\hat{v}_s^j, v_R\right)}{2}M_j\right\|_2 \\ L_{\text{Smooth}} = \sum_{i=1}^{P_N}\dfrac{1}{P_N}\left(e^{-\alpha_1|\nabla v_R|}\left|\nabla \boldsymbol{D}_i\right| + e^{-\alpha_2|\nabla^2 v_R|}\left|\nabla^2 \boldsymbol{D}_i\right|\right) \\ L_{FV} = \gamma_1 F_{C,1/2} + \gamma_2 F_{C,1/4} + \gamma_3 F_{C,1/8} \end{cases} \tag{4}$$

式中：前 3 项称作逐像素损失 L_P，分别表示光度一致性损失、结构相似度损失和平滑损失；m_j 表示掩码 M_j 中有效点的总数；∇表示梯度；P_N 表示参考视

点 v_R 中点的总数；其他为权重的常数。本文特征损失利用视点输入的 CNN 3 个不同层的公共特征体加权得到，3 层的大小分别是原始输入的多视点图像大小的二分之一、四分之一和八分之一。上述各损失可以最终帮助输出参考视点中的完整深度图。使用 3D U-Net 正则化，经由 Softmax 函数处理从而得到概率体积(Probability volume)，最后可通过加权求和获得深度图。利用投影坐标变换等方式，基于估计的深度信息，将二维图像重建为三维模型。

2.2　基于扩散模型优化的神经辐射场网络

在本系统中，设计了一个单独的深度渲染优化模块，用于感知场景光线坐标和指向向量优化对三维点的 RGB 值和密度的估计。由于 NeRF 是按场景优化的隐式表示，系统引入了基于扩散模型(Diffusion model)的函数提取形状和颜色的先验信息，构建底层一致的三维场景来指导其内容学习，实现更高效、更精确的三维空间感知。融合在数据校准过程中获得的相机参数，使用细化后的 RGB 值和密度，通过基于 NeRF 的体渲染方法，可以进一步优化在上一节输出的深度并指导三维重建。

扩散模型通过迭代去噪对数据分布 $p_{\text{data}}(x)$进行建模，通过匹配进行训练。给定样本 $x\sim p_{\text{data}}$ 和噪声分布 $\epsilon\sim N(0,\boldsymbol{I})$，用参数化 θ 建立去噪器 ϵ_θ，接收扩散输入 $x_t(\epsilon,t,x)$，并通过最小化去噪分数进行迭代，直到匹配学习目标。记输入的多视点数据包含的稀疏特征数据为 x_i，算法根据上述基本模型，结合去噪扩散隐式模型(Denoising Diffusion Implicit Models, DDIM)方法[29]，建立本系统需要的扩散模型函数，目标是紧密逼近空间中光线物理信息的真实分布模型 θ 所对应的数据分布 $p(x)$。其确定性前向过程对应的神经网络逐渐向样本添加噪声，以满足

$$q\left(x_t \mid x_{t-1}\right)=N\left(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t\boldsymbol{I}\right) \tag{5}$$

式中：β_t 表示时间上第 t 个方差表。然后，扩散模型函数使用神经网络学习相反的去噪过程，即对应空间信息的感知。去噪过程同样使用高斯分布，即

$$p_\theta(x_{t-1} \mid x_t)=N\left(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)\right) \tag{6}$$

式中：μ_θ 和 Σ_θ 分别表示均值和方差。进一步，扩散模型函数通过利用预训练变分自动编码器的语义信息提升效率。神经网络经过优化，可以根据图像和文本指令调节输入来预测呈现的噪声。潜在的语义扩散目标为

$$L(\boldsymbol{C}_i,\boldsymbol{D}_i)=\mathbb{E}_{\varepsilon\sim N(0,1),t}\left[\left\|\omega-D_\theta\left(z_t,t,\varepsilon(\boldsymbol{C}_I),\boldsymbol{C}_T\right)\right\|_2^2\right] \tag{7}$$

式中：$\varepsilon(\cdot)$表示编码器；z_t 表示时间 t 处的噪声输入；$\boldsymbol{C}_I$ 表示条件图像，对应需要重建的数据；$\boldsymbol{C}_T$ 表示空间语义输入；ω 表示时间 t 的预测噪声。在扩散模型网络训练完成后，可以作为函数使用给定时间的噪声输入和预测噪声导出估计的潜在的扩散值，表征空间中光线约束下的颜色密度和体积密度，记为$(\boldsymbol{C}_i,\boldsymbol{D}_i)$，优化 NeRF 的学习。

给定三维点 $\boldsymbol{x}$ 和特定的观看方向 $\boldsymbol{d}$，NeRF 学习一个隐式函数 f，该函数估计空间光线的微分密度 σ 和 RGB 颜色 $\boldsymbol{c}$，记为 $f(\boldsymbol{x},\boldsymbol{d})=(\sigma,\boldsymbol{c})$。对于所采集数据中的任意视点 v_n，光线 $\boldsymbol{R}_i(r)$通过相机中心 $\boldsymbol{o}$ 所产生的任何像素 p_i 和三维点(x_i,y_i,z_i)可以定义为 $\boldsymbol{R}_i(r)=\boldsymbol{o}+r\boldsymbol{d}_i$，其中 r 为光线上的位置参数，$\boldsymbol{d}_i$ 为由观看方向 θ_i 和 ϕ_i 表示的视线矢量。使用以 RGB 为格式的颜色值 $c_i(r)=(\boldsymbol{R}_i(r),\boldsymbol{d}_i)$，射线上三维点的密度 $\sigma_i(r)=(\boldsymbol{R}_i(r))$即表示不透明度。通过采样光线 $\boldsymbol{R}_i(r)$上的任意 N 个点 $\boldsymbol{r}'$，基于采样的黎曼和，像素 p_i 的 RGB 值将来自任何对象入射辐射及其深度定义为

$$\hat{\boldsymbol{C}}_i(\boldsymbol{r}')=\sum_{j=1}^{N}T_j\left(1-\exp\left(-\sigma_j\delta_j\right)\right)\boldsymbol{c}_j \tag{8}$$

$$\hat{\boldsymbol{D}}_i(\boldsymbol{r}')=\sum_{j=1}^{N}T_j\left(1-\exp\left(-\sigma_j\delta_j\right)\right)r_j \tag{9}$$

式中：T_j 表示累积透射率，辅助算法实现空间中遮挡部分的生成，满足 $T_j=\exp\left(-\sum_{k=1}^{j-1}\sigma_k\delta_k\right)$；且 $\delta_j=r_{j+1}-r_j$ 为位于光线上的相邻采样点之间的距离。

借助扩散模型学习光线分布过程中得到的有关参数和无监督 MVS 网络的估计深度值，NeRF 进一步感知需要重建的数据在空间上的分布，得到所估计深度的优化值。在网络迭代过程中，其损失函数由 2 部分组成：其一是颜色一致性损失，表征生成过程中，由特定的采集设备参数 $\boldsymbol{P}=(K;[\boldsymbol{R},t_v])$ 产生的射线集 $\boldsymbol{R}$ 返回的颜色和深度构建损失，即

$$\begin{cases}\arg\min L_{\text{NeRF}}=L_{\text{color}}+\lambda_d L_{\text{depth}}\\ \text{s.t. } L_{\text{color}}=\mathbb{E}_{r\in\boldsymbol{R}(\boldsymbol{P})}\left\|\hat{\boldsymbol{C}}_i(r)-\boldsymbol{C}_i(r)\right\|^2\\ L_{\text{depth}}=\mathbb{E}_{r\in\boldsymbol{R}(\boldsymbol{P})}\left\|\hat{\boldsymbol{D}}_i(r)-\boldsymbol{D}_i(r)\right\|^2\end{cases} \tag{10}$$

式中：$\boldsymbol{C}_i(r)$和 $\boldsymbol{D}_i(r)$分别表示所生成的参考视点的像素颜色和深度值；λ_d 表示在颜色损失和深度损失之间平衡的权重参数。在实际计算过程中，深度损失仅针对穿过具有深度图的像素的光线进行计算，而不会对所有像素进行遍历，这有利于降低计算的负载，从而拓展可用于生成的内容范围。在本系统中，基于 NeRF 的方法使用的多层感知(Multi-Layer Perception, MLP)迭代优化无监督网络的输出深度

图，计算损失 L_{NeRF} 并在给定次数的迭代之后选取具有最小损失的情况下输出生成结果。在实现过程中，MLP 均由每个输出 256 维特征向量的 8 个全连接层组成。

通过在体渲染(Volume rendering)中累积对应光线上的三维点及其密度来映射深度估计结果。引入扩散模型学习到的空间先验信息，当系统通过不同数据集实现三维重建时，不需要将所有视点数据输入到 NeRF 中，在运行一组数据后，所获得的 NeRF 模型可以直接在其他数据上使用，从而降低了计算复杂度和推理时间。

3 实验对比

本系统对应的神经网络可在云服务器上进行实现。云服务器配备 Intel(R) Xeon(R) Silver 4216 CPU @ 2.10 GHz、4 块 NVIDIA GeForce RTX3090 GPU 以及 10 TB 存储空间。在训练设置方面，训练 15 个轮次(Epoch)收敛，每个轮次耗时 8 h。该时间包括了无监督 MVS 网络和 NeRF 优化模块并行迭代的总时长。通过像素重投影生成了点云模型实现结果的可视化。

3.1 DTU 数据集

DTU 数据集[30]是在实验室可控条件下、相机轨迹固定采集的室内数据集。其通常作为评估 MVS 网络相关方法的基准数据集，由多组多视点图像和相应的相机参数组成。包含 124 组数据，每组数据有 49 或 64 个视点，总共有 27 097 个训练样本，涵盖 7 种不同光照条件，并被划分为 79 个训练集、18 个验证集和 22 个测试集。在基于 DTU 数据的测试中，本模型在训练集上进行训练优化，在测试集上进行实验评价。

本网络使用 PyTorch 在 DTU 数据集上进行训练，采用一个三阶段的多尺度流水线(Pipeline)来构建无监督网络。对于每个阶段，使用不同的特征图和 3D-CNN 网络参数。数据集中的输入图像首先被调整为 600×800 大小，然后裁剪成 512×640 的图像块。细化后的真实深度图的分辨率设置为 160×128，实现从 425 mm 到 935 mm 尺寸的均匀采样。为了平衡损失函数中的不同权重，设置相关的常量 λ_1=0.8，λ_2=0.2，λ_3=0.0067，λ_4=1.0，λ_d=0.1，β_1=0.5，β_2=0.5，α_1=0.5 及 α_2=0.5。此外，设置 γ_1=0.2，γ_2=0.8 和 γ_3=0.4。整个网络使用 Adam 优化器根据梯度进行 15 个轮次的优化，初始学习率为 1e-4，在第 10，12 和 14 个轮次后学习率缩小为原来的一半。在云服务器的 GPU 上以 4 的批量(Batch size)大小进行训练，每个轮次迭代 50 000 回(Iteration)。在每回迭代中，使用 DTU 数据集中的 1 张参考图像和 3 张源图像。在每一轮次后，无监督深度学习网络的输出也被同步发送到 NeRF 优化模块。NeRF 网络以与无监督网络当轮相同的学习率感知训练集数据，在 15 轮中独立迭代并选择损失最小的输出作为最终的三维重建结果。

为确定输入网络的图像数量 N，首先进行了参数优化实验。整个 DTU 数据集的定量结果通过官方 MATLAB 评估代码[30]计算得出。在定量比较上采用 3 个官方指标来评估性能(表 1)。所有指标的值越小，相应方案越好，其中准确率(Accuracy)和完整性(Completeness)是 2 个表示重建质量的绝对距离指标，总体指标(Overall)是上述 2 个指标的平均值。

表 1 DTU 数据集上的参数优化实验

Table 1 Parameter Optimization on DTU dataset

参数	准确率	完整性	总体指标
N=3	0.352	0.276	0.314
N=4	0.338	0.256	0.297
N=5	**0.337**	**0.256**	**0.295**
N=6	0.340	0.261	0.300
N=7	0.357	0.284	0.321

注：加粗数据表示最优值。

从表 1 可见，由于数据集覆盖的内容范围有限，过小的参数可能会损失精度，而过大的参数则会导致类似过拟合的现象。参数设置为 N=5 时能够获得更精细的深度图，图 2 展示了本系统基于深度估计后重建的三维点云可视化结果。从平均指标上看，误差<2 mm，<4 mm 和<8 mm 的点比例分别为 0.895，0.927 和 0.956，显示出本系统在提升深度估计方面的有效性。

基于 DTU 测试集，验证了本文提出的一系列神经网络的深度预测性能，将输入图像分辨率设置为 1 152×1 600。各方案的定量实验结果见表 2，本模型的总体得分相较其他方案平均提高了 24.6%，显著超过了传统方法[10]和有监督学习方法[2-3]。ColNeRF[26]将与本文提出的类似的 NeRF 优化模块引入到 Colmap 重建结果中，虽然提高了重建模型的准确率，但在完整性方面表现不佳。通过利用无监督学习网络，本系统较其整体性能提高了 17.6%。此外，本系统也优于直接通过 NeRF 构建损失函数的无监督学习的 RC-MVSNet[20]，准确率较其提高了 11.1%，并以较低的训练成本在其他指标上也取得了更好的成绩。完整性从 CL-MVSNet[5]

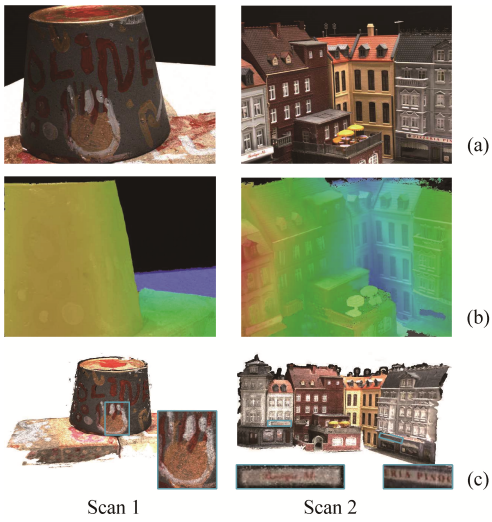

图 2　DTU 数据集上的点云重建结果((a) 原图；(b) 深度结果；(c) 重建结果)

Fig. 2　Point cloud results on DTU dataset ((a) Original data; (b) Depth estimation result; (c) Reconstruction result)

表 2　DTU 数据集上的重建性能

Table 2　Evaluation metrics on DTU dataset

方案	准确率	完整性	总体指标
Colmap[10]	0.400	0.664	0.532
MVSNet[2]	0.396	0.527	0.462
M3VSNet[3]	0.636	0.531	0.583
Unsup-MVS[17]	0.881	1.073	0.977
RC-MVSNet[20]	0.396	0.295	0.345
CL-MVSNet[5]	0.375	0.283	0.329
RA-MVSNet[31]	**0.326**	0.268	0.297
CT-MVSNet[19]	0.341	0.264	0.302
ColNeRF[26]	0.384	0.378	0.381
本文方案	0.337	**0.256**	**0.295**

注：加粗数据表示最优值。

的 0.283，基于 Transformer 的 CT-MVSNet[19]的 0.264 提高到本系统的 0.256，证明了本网络在解决不完整预测方面的有效性。特别是在准确率保持相似性能的同时，完整性和总体指标优于空间分层感知的 RA-MVSNet[31]。

进一步进行了模块间的消融实验，展示了本系统不同组件在三维重建性能上的提升。实验包含以下 3 种情况：①仅使用带有像素损失的无监督深度学习网络，记为 L_P；②仅使用带有像素和特征损失的无监督深度学习网络，记为 L_P+L_{FV}；③使用基于 NeRF 优化的无监督深度学习网络，即本文所提出的完整的模型，记为 L and L_{NeRF}。定量结果的比较见表 3，完整的模型评估结果具有最高的准确率，证实了本算法各组件对于重建结果均是不可或缺的。

为了更直观地展示本方案的消融实验结果，在图 3 中提供了使用不同核心组件组合训练的模型的定性比较。可以看到，本文无监督 MVS 网络以高精度恢复物体表面的内容，NeRF 对其进行改进，以高精度估算出物体边界的深度，得到更加完整的点云的重建结果。本系统的输出与原始真实场景的视觉外观相似，重建结果中很好地保留了表面高光和标志上的文字等视觉效果。

表 3　DTU 数据集上的消融实验

Table 3　Ablation study on DTU dataset

方案	准确率	完整性	总体指标
L_P	0.432	0.349	0.391
L_P+L_{FV}	0.391	0.285	0.338
L and L_{NeRF}	**0.337**	**0.256**	**0.295**

注：加粗数据表示最优值。

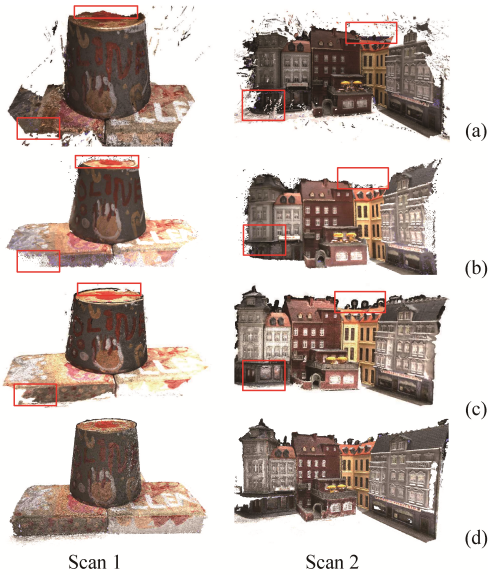

图 3　DTU 数据集上的消融实验可视化结果((a) L_p；(b) L_p+ L_{FV}；(c) L 和 L_{NeRF}；(d) 数据集参考结果)

Fig. 3　Visual results of ablation study on DTU dataset ((a) L_p; (b) L_p+ L_{FV}; (c) L and L_{NeRF}; (d) Dataset baseline result)

3.2　Tanks and Temples 数据集

为了评估本系统的可行性和泛化能力，采用宽基线的 Tanks and Temples 数据集[32]作为测试集。基于大规模三维重建的数据集，包含 14 个场景，包括像“坦克(Tank)”和“火车(Train)”这样的建筑规模的单个物体。在这个基准测试中，为生成的点云计算 F 分数(F-score)，并将不同场景下的分数与上述最先进的方案进行比较。

在 DTU 数据集上的模型在未进行任何微调的情况下迁移到 Tanks and Temples 数据集上，考虑到数据的规模将图像尺寸设置为 1 920×1 056，输入图像数量 N=7。为获得更好的测试结果，在实验中将本文提出的 MVS 网络使用光度约束对采集场景的预测深度图进行过滤。光度约束用于衡量多视点间的匹配质量，网络预测的置信度值较低的深度将

被舍弃。具体而言，设置估计深度概率低于 0.3 的像素将被深度图丢弃。此外，系统通过将数据迭代输入到预训练模型中获得结果，迭代 10 000 次且不进行训练，所有深度值由迭代过程中损失最小的网络输出进行融合和处理。基于 NeRF 的几何约束也用于衡量多视点深度一致性，与相邻视点深度不一致的深度也将被丢弃。最后将多视点数据转换为点云评估相关结果。

表 4 展示了在 Tanks and Temples 数据集上的代表性评估结果。本系统在对比方法中取得了最佳性能，可证实本方法在大规模场景中的有效性。特别地，本系统平均得分比 M3VSNet[3]显著提高了 18.61 分，比 RC-MVSNet[20]高了 6.78 分；与 ColNeRF[26]中的模块化组合或引入 Transformer 网络的 CT-MVSNet[19]相比，无监督 MVS 视觉网络与 NeRF 的组合能获得更多优势，平均提高了 3.47 分。此外，还比较了算法在深度假设参数上的取值，表征深度估计范围。可以看到，深度范围在 D=160 以内时性能最优，当范围过小时算法无法有效呈现对于大规模场景的全部信息，而继续增大取值空间则引入了一定的生成噪声，性能反而略有下降。

表 4 Tanks and Temples 数据集上的重建性能

Table 4 Evaluation metrics on Tanks and Temples dataset

方案	Lighthouse	Panther	Train
Colmap[10]	56.43	46.97	42.04
MVSNet[2]	50.79	50.86	34.69
M3VSNet[3]	44.42	44.95	30.31
Unsup-MVS[17]	42.03	44.00	36.45
RC-MVSNet[20]	53.49	52.30	49.37
CL-MVSNet[5]	60.02	59.97	52.28
RA-MVSNet[31]	64.78	65.60	58.08
CT-MVSNet[19]	62.60	64.83	58.68
ColNeRF[26]	60.23	59.46	52.57
本文方案			
D=128	61.17	61.20	53.14
D=160	**64.97**	**65.90**	**58.74**
D=192	64.89	65.85	58.71

注：加粗数据表示最优值；数值越大越好。

图 4 所示的可视化结果更直观展示了本系统在 3 个示例场景中的三维重建能力。相比对比的基线方案中更多地去调整无监督学习网络的损失函数，希望以一个综合的神经网络实现性能优化，本文提出的三维重建方案引入了 NeRF 隐式表达，额外引入空间中语义的感知情况，对无监督网络的结果进行了补偿，实现在重建完整度、离散点误差等指标表征的重建高精度泛化性能。

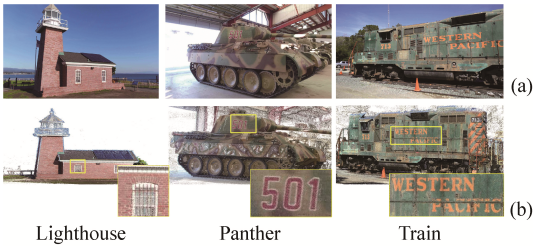

图 4 Tanks and Temples 数据集上的点云重建结果
((a) 原图；(b) 重建结果)

Fig. 4 Point cloud results on Tanks and Temples
((a) Original data; (b) Reconstruction result)

3.3 NERULN 数据集

从自然场景采集的 NERULN 数据[33]被输入到本文提出的无监督深度学习网络以及最相关联的 M3VSNet[3]，RC-MVSNet[20]和 ColNeRF[26]对比方案实现重建，以比较生成的三维点云。具体过程与上一小节类似，直接将数据输入到预训练网络模型中，相应地将视点分辨率调整为 1 280×960，输入数量为 N=5。在预测深度值的过程中，估计深度低于 0.25 的像素将被舍弃。迭代完成后，评估重建的精细度、总执行时间和所需的系统资源。

为了说明定性比较情况，图 5 展示了同一帧重建点云的视觉结果。基于参考/源视点的 M3VSNet[3]

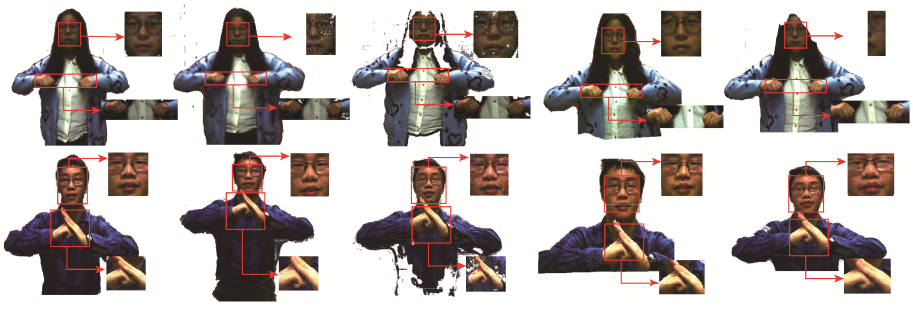

图 5 NERULN 数据集上的点云重建结果((a) 本文方案(完整架构)；(b) 本文消融方案(只有无监督 MVS 网络)；(c) Co1NeRF；(d) RC-MVSNet；(e) M3VSNet)

Fig. 5 Point cloud results on NERULN dataset ((a) Proposed system (Full); (b) Ablation proposed system (L only); (c) Co1NeRF; (d) RC-MVSNet; (e) M3VSNet)

和 RC-MVSNet[20]在所采集的大场景数据中无法重建完整的三维模型。例如，其无法恢复人体肘部和腰部的视觉信息，因为参考视点本身未覆盖到这些内容。相反，本系统能够重建出更完整的结果，因为基于 NeRF 优化的算法可以将肘部和腰部的视觉信息额外合成进重建的模型中。此外，观察本系统的模型高频部分，可以看到生成的伪影或空洞比采用类似 NeRF 优化方法的 ColNeRF[26]更少，三维模型在脸部边缘和手腕处更平滑，并且具有更完整的纹理细节。

各方案的性能统计比较见表 5，包含了只有无监督 MVS 网络(L only)和完整架构(Full)的消融实验结果。本文所输出的模型具有相对较多的点和面，较多的重建细节意味着所获得的点云模型具有更高的完整性。此外利用三维点从相邻像素重投影的颜色误差，本文输出完整模型的平均重投影误差约为 0.168 px，而 M3VSNet[3]输出的模型重投影误差约为 0.284 px。与 M3VSNet 相比，本系统将误差降低了 40.8%。统计数据还表明，本系统在应对采集到光场数据的三维重建方面至少比基于 NeRF 的 RC-MVSNet[20]性能高出 11.1%。虽然本系统对总共 100 个视点的处理时间平均为 321.3 s，比 RC-MVSNet 长约 8.3%，但在保持更好重建性能的同时承担略大一点的时延仍是可接受的。本系统比 ColNeRF[26]取得了更全面的重建结果，比其 0.204 px 的误差低了 17.6%。

表 5　NERULN 数据集上的重建性能

Table 5　Evaluation metrics on NERULN dataset

方案	点数量	面数量	重建误差/px	处理时间/s	模型规模/MB
M3VSNet[3]	519 291	103 854	0.284	354.9	6320
RC-MVSNet[20]	504 690	100 574	0.189	294.5	9189
ColNeRF[26]	538 372	119 578	0.204	341.5	5964
本文-L only	530 249	118 988	0.202	**286.7**	5970
本文-Full	**560 560**	**130 609**	**0.168**	321.3	8672

注：加粗数据表示最优值；点面数量越大越好，误差和处理时间越小越好。

本系统是将模块化的立体视觉解决方案引入重建服务的一次有潜力的尝试。无监督网络与 NeRF 解耦，分别承担特征提取与体渲染职责，构建了从粗到细的特征感知和三维重建体系。通过引入场景包含的光线物理约束，系统具有良好的泛化能力，能够适应窄基线(如 DTU 数据集)和宽基线(如 Tanks and Temples 数据集、NERULN 数据集)等不同情况。基于视觉的解决方案对采集到的物理世界场景没有限制，几乎可以应用于重建出现在相机视野内的所有物体。本系统模型规模相较传统方案[3,26]要更大，但由于模块间的独立性，三维重建工作可以更加灵活地进行，如在云端进行 MVS 网络进行基本重建，再通过边缘设备上的 NeRF 获得精确结果。

然而，目前本系统的运行时间不足以支持实时处理。与网络通信延迟相比，计算时间仍然是主要的性能瓶颈。这可能需要采用轻量级算法来减少整体处理延迟。在重建任务的效率和准确性之间进行权衡至关重要。考虑到在自然场景中采集的大量视角以及系统需要按顺序处理每个视角的特点，可以采用数据压缩来减少整体运行时间。通过设计特定的边云协同框架，基于几何-语义解耦或动态自适应权重分配等损失函数机制，可以进一步降低传输的带宽消耗以及基于无监督学习的深度预测的计算复杂度。

4　结束语

本文提出了一种新颖的 MVS 视觉系统，基于改进的视点光度一致性建立无监督神经网络来高效生成深度图。通过在一个独立模块中集成基于扩散模型的 NeRF 渲染和深度值细化，获得了细粒度的三维表示。在 DTU 和 Tanks and Temples 数据集上与其他基于学习的方法相比，本系统取得了具有竞争力的深度结果，显示出较强的有效性和鲁棒性。此外，在自然场景宽基线图像数据集中进行的泛化测试，证明了本方法在自然场景中的可扩展性，表明确能够很好地重建出各种具有挑战性的真实世界三维模型。未来的工作可以关注于开发一种用于无监督 MVS 视觉的轻量级方案，利用移动边缘计算等架构进一步提高系统的性能指标。

参考文献 (References)

[1] LEE L H, BRAUD T, ZHOU P Y, et al. All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research agenda[J]. Foundations and Trends® in Human-Computer Interaction, 2024, 18(2/3): 100-337.

[2] YAO Y, LUO Z X, LI S W, et al. MVSNet: depth inference for unstructured multi-view stereo[C]//The 15th European Conference on Computer Vision – ECCV 2018. Cham: Springer, 2018: 785-801.

[3] HUANG B C, YI H W, HUANG C, et al. M3VSNET: unsupervised multi-metric multi-view stereo network[C]//2021 IEEE International Conference on Image Processing. New York: IEEE Press, 2021: 3163-3167.

[4] LI J L, LU Z D, WANG Y Q, et al. DS-MVSNet: unsupervised

multi-view stereo via depth synthesis[C]//The 30th ACM International Conference on Multimedia. New York: ACM, 2022: 5593-5601.

[5] XIONG K Q, PENG R, ZHANG Z, et al. CL-MVSNet: unsupervised multi-view stereo with dual-level contrastive learning[C]//2023 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2023: 3746-3757.

[6] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[C]//The 16th European Conference on Computer Vision. Cham: Springer, 2020: 405-421.

[7] 王道累, 丁子健, 杨君, 等. 基于体素网格特征的 NeRF 大场景重建方法[J]. 图学学报, 2025, 46(3): 502-509.
WANG D L, DING Z J, YANG J, et al. Large scene reconstruction method based on voxel grid feature of NeRF[J]. Journal of Graphics, 2025, 46(3): 502-509 (in Chinese).

[8] ZAWISH M, DHAREJO F A, KHOWAJA S A, et al. AI and 6G into the Metaverse: fundamentals, challenges and future research trends[J]. IEEE Open Journal of the Communications Society, 2024, 5: 730-778.

[9] 刘鑫, 李洋, 冯胜杰, 等. 面向 RGB-D 数据的特征线提取和表示算法[J]. 图学学报, 2025, 46(3): 542-550.
LIU X, LI Y, FENG S J, et al. Line extraction and representation algorithm for RGB-D data[J]. Journal of Graphics, 2025, 46(3): 542-550 (in Chinese).

[10] SCHÖNBERGER J L, ZHENG E L, FRAHM J M, et al. Pixelwise view selection for unstructured multi-view stereo[C]//The 14th European Conference on Computer Vision. Cham: Springer, 2016: 501-518.

[11] HEEP M, ZELL E. ShadowPatch: shadow based segmentation for reliable depth discontinuities in photometric stereo[J]. Computer Graphics Forum, 2022, 41(7): 635-646.

[12] LIANG J, WANG R J, PENG R, et al. High fidelity aggregated planar prior assisted PatchMatch multi-view stereo[C]//The 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 3141-3150.

[13] TANG J Y, CAI Y G, GAO X S, et al. Generalized sampling of non-local textural clues multi-view stereo framework[C]//The 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 11222-11225.

[14] XU H B, CHEN W T, SUN B G, et al. RobustMVS: single domain generalized deep multi-view stereo[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10): 9181-9194.

[15] ZHU J, PENG B, LIU B Z, et al. Self-constructing stereo correspondences for unsupervised multi-view stereo[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10732-10742.

[16] JIANG J F, CAO M F, YI J, et al. DI-MVS: learning efficient multi-view stereo with depth-aware iterations[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing New York: IEEE Press, 2024: 3180-3184.

[17] KHOT T, AGRAWAL S, TULSIANI S, et al. Learning unsupervised multi-view stereopsis via robust photometric consistency[EB/OL]. (2019-06-06)[2025-01-27]. https://arxiv.org/abs/1905.02706.

[18] RENDLE G, KRESKOWSKI A, FROEHLICH B. Volumetric avatar reconstruction with spatio-temporally offset RGBD cameras[C]//2023 IEEE Conference Virtual Reality and 3D User Interfaces. New York: IEEE Press, 2023: 72-82.

[19] WANG S C, JIANG H, XIANG L. CT-MVSNet: efficient multi-view stereo with cross-scale transformer[C]//The 30th International Conference on Multimedia Modeling. Cham: Springer, 2024: 394-408.

[20] CHANG D, BOŽIČ A, ZHANG T, et al. RC-MVSNet: unsupervised multi-view stereo with neural rendering[C]//The 17th European Conference on Computer Vision. Cham: Springer, 2022: 665-680.

[21] DENG K L, LIU A, ZHU J Y, et al. Depth-supervised NeRF: fewer views and faster training for free[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 12872-12881.

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

[23] SANTO H, OKURA F, MATSUSHITA Y. MVCPS-NeuS: multi-view constrained photometric stereo for neural surface reconstruction[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2024: 20475-20484.

[24] ZHU D X, KONG H R, QIU Q, et al. Multi-view stereo network based on attention mechanism and neural volume rendering[J]. Electronics, 2023, 12(22): 4603.

[25] WEI Y, LIU S H, ZHOU J, et al. Depth-guided optimization of neural radiance fields for indoor multi-view stereo[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10835-10849.

[26] ITO S, MIURA K, ITO K, et al. Neural radiance field-inspired depth map refinement for accurate multi-view stereo[J]. Journal of Imaging, 2024, 10(3): 68.

[27] ZHU H X, CHEN Z B. CMC: few-shot novel view synthesis via cross-view multiplane consistency[C]//2024 IEEE Conference Virtual Reality and 3D User Interfaces. New York: IEEE Press, 2024: 960-968.

[28] SCHÖNBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 4104-4113.

[29] CAO T S, KREIS K, FIDLER S, et al. TexFusion: synthesizing 3D textures with text-guided image diffusion models[C]//2023 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2023: 4146-4158.

[30] AANÆS H, JENSEN R R, VOGIATZIS G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

[31] ZHANG Y S, ZHU J K, LIN L X. Multi-view stereo representation revist: region-aware MVSNet[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2023: 17376-17385.

[32] KNAPITSCH A, PARK J, ZHOU Q Y, et al. Tanks and temples: benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics, 2017, 36(4): 78.

[33] PAN Y X, LIU Y, ZHANG L. LiTrix: a lightweight live light field video scheme for metaverse stereoscopic applications[J]. IEEE Internet of Things Magazine, 2023, 6(2): 137-142.