特殊表面检测和分割使用 SAM&CLIP

摘要

本文探讨了多视图立体视觉重建（MVS）在城市环境中图像对齐和优化的关键挑战，重点讨论了传统多视角重建方法中因特殊表面材质（如玻璃幕墙、窗户和金属结构）带来的一致性求解和像素匹配的困难。提出有效的图像对齐技术以提升模型精度、鲁棒性和效率。介绍了采用块对块匹配策略以及识别并分割图像中的特殊表面的必要性，由于现有预训练网络的局限性和数据集构建的高成本，文中提出使用分割一切模型（SAM）进行高效图像分割，并结合 CLIP 多模态模型对 SAM 输出进行分析，以精确分割特定玻璃表面。考虑到这些模型的广泛泛化能力，该策略预期将优化 MVS 重建中的多视角对齐过程，并显著提高城市场景 MVS 重建的质量，为 MVS 重建领域开辟新的研究方向。

关键词：语义分割；三维重建；多模态

1 引言

在多视图立体视觉重建（MVS, Multi-view Stereo）的研究中，针对城市环境的图像对齐和优化一直是一个关键且复杂的挑战。在传统的多视角重建方法中，主要依赖于多个视点之间图像像素的密集匹配来计算空间点云，进而构建出三维网格模型。然而，复杂的城市场景中包含了多种特殊的表面材质，例如反光的玻璃幕墙、窗户以及金属结构，这些材质对于多视点间的一致性求解和像素之间的密集匹配带来了显著的挑战。

为了克服这些挑战，有效的图像对齐技术显得尤为重要。这不仅能提升重建模型的准确性，而且能增强算法的鲁棒性和运行效率。通过减少误匹配的发生，可以避免在后续处理阶段错误的传播，从而显著提高重建质量。此外，高质量的图像对齐还为下游任务，例如纹理映射和模型优化，提供了一个更为稳定的基础。在这一背景下，采用块对块（Patch-to-Patch）的匹配策略成为一种可行的解决方案，这种策略能够专门处理那些具有挑战性的表面特征，并减轻复杂光学属性的影响。

实现这种匹配策略的第一步是有效地识别并分割图像中的特殊表面材质。当前，许多预训练的神经网络大多基于特定的有监督数据集进行训练，但它们可能不适用于 MVS 中复杂的城市场景。自行构建和标注数据集虽然可以提供更准确的检测结果，但这过程耗时且成本高昂。

鉴于此，我们提出使用分割一切模型（SAM, Segment Anything Model）[11] 进行高效的图像分割。此外，我们进一步建议结合使用 CLIP 或其他多模态模型对 SAM 的输出进行深入分析，以准确判断分割出的区域是否为特定的玻璃表面。考虑到这些大型模型在广泛的应用场景中展现出了卓越的泛化能力，这种策略有望在 MVS 重建中优化多视角对齐过程，从

而显著提高城市场景中的 MVS 重建质量。通过这种方法，我们不仅能够更精准地处理复杂城市场景中的特殊表面，还能够为 MVS 重建领域提供一个新的研究方向。

2 相关工作

2.1 多视角重建

多视角立体几何（MVS）作为计算机视觉研究的重要领域，近年来已取得显著的进展。这一领域的发展得益于运动结构（SfM, Structure from Motion）算法的突破和公共基准测试的普及 [3,20]，过去十年中涌现出众多卓越的 MVS 方法。这些方法主要依靠不同视角下图像之间的密集像素匹配进行三维重建 [4]。然而，这种方法在处理低纹理、重复纹理或反射表面等特殊情况时，往往无法实现准确匹配，导致重建结果出现误差。

针对这一挑战，研究正在探索更为高级的解决方案。例如，研究提出结合深度学习和传统的几何方法，以提高对复杂表面的理解和处理能力 [7,9,24]。这些方法可能包括利用先进的深度学习网络来进行特征的匹配，或者结合多种传感器数据来增强对复杂场景的理解。此外，为了进一步提升重建的准确性，研究者也在探索自适应算法和先验知识的整合应用，以便更好地处理 MVS 重建中多样化的表面特性。

针对这一问题，研究者提出了基于块匹配（Patch Match）的方法 [1,9,17–19]。块匹配算法通过估计每个像素对应的物体表面平面来进行深度估计，将问题转化为平面估计问题。算法首先对像素的初始平面假设进行随机初始化，然后根据光度一致性原则计算这些平面假设的置信度。最终，算法将那些置信度高的平面假设传播至相邻像素，从而高效准确地估计出每个像素的深度。

尽管基于块匹配的方法在处理低纹理表面时展现出一定的优势，但在处理复杂城市场景中频繁出现的玻璃表面时，这种方法却显得力不从心。这是因为玻璃表面的光学特性，如反射和透明性，使得基于块匹配的方法难以有效捕捉和处理这些表面的深度信息。此外，玻璃表面的存在不仅影响了单个像素的深度估计，还可能导致整个场景的几何结构重建不准确。总体而言，虽然 MVS 领域已经取得了显著的进步，但在处理城市场景中的特殊表面，如玻璃，时仍面临着重大挑战。未来的研究需要更加深入地理解和解决这些挑战，以推动 MVS 技术在更广泛、更复杂应用场景中的发展。

2.2 玻璃和镜面分割

特殊表面检测在计算机视觉领域中扮演着一项基础而关键的任务。尤其在三维重建的复杂场景中，如城市环境，玻璃表面的普遍存在增加了检测的复杂度。这一挑战主要源于玻璃表面的光学特性，如折射和反射，使得它们不具有一致的外观特征。相比于通用对象检测和分割，这些特殊表面的检测工作更为复杂。常见的视觉任务往往会在处理玻璃表面时提供错误的信息，比如将玻璃后的物体误识别为玻璃表面，或者在深度预测时将反射或折射的物体深度误判为玻璃本身的深度。

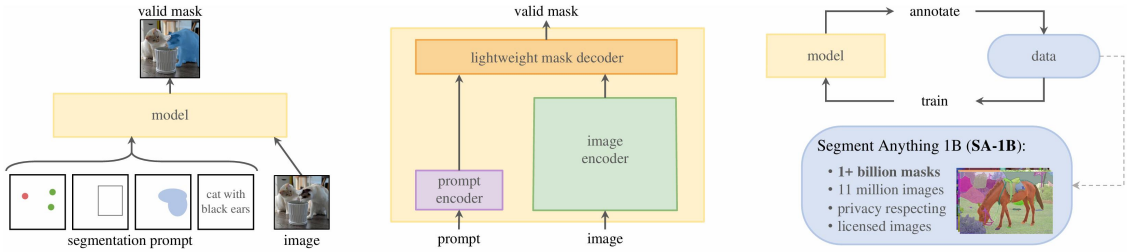
为了应对这些挑战，学者们已经提出了多种解决方案。Yang [23] 等提出了第一个大规模特殊表面检测数据集和第一个深度学习模型 MirrorNet，通过学习镜面内部和外部之间的上下文对比度进行镜面的检测。继此之后，许多研究者通过设计创新的网络结构和应用策略来

逐步提升对各种特殊表面的检测能力 [2, 5, 6, 8, 10, 13–15, 21–23, 25, 26]，同时也有学者在最近提出使用自监督方式来检测环境中的特殊表面 [12]。

尽管这些研究在小场景级别的玻璃或镜面检测上取得了显著成效，但在应对更为复杂的城市场景时，它们往往表现不佳。这一问题部分源于训练数据集分布的局限性，导致这些模型无法直接应用于大场景图像中的特殊表面检测。因此，针对城市环境的特殊表面检测仍然是一个待解决的课题，需要更多地考虑场景复杂性和数据集多样性，以及如何有效整合这些因素以提高检测算法的泛化能力和准确性。

3 本文方法

3.1 本文方法概述



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

图 1. 本文的研究目标是开发一个基础的图像分割模型，该模型在结构上由三个相互依赖的组件构成：(a) 一个能够适应灵活提示的分割任务框架。(b) 一个集成了提示工程技术的分割模型（SAM），该模型不仅能够辅助数据标注过程，还能在 0 样本或少样本的情况下实现对各种任务的迁移。(c) 一个高效的数据收集引擎，专门用于构建 SA-1B 数据集，该数据集包含超过 1100 万张图像以及超过 11 亿个用于训练 SAM 模型的图像掩码。

该论文提出了一个图像分割的基础模型。该项目通过一个高效的模型在数据收集循环中工作，建立了一个前所未有的大规模分割数据集，包含超过 11 亿个掩码和 1100 万张图像。该模型经过设计和训练，可以响应多种提示（Prompt），使其能够在零样本的情况下迁移到新的图像分布和任务中。后期对模型在多个任务上的能力进行评估时，发现其零样本性能不仅令人印象深刻，而且在很多情况下能够与之前的全监督方法相媲美，甚至更胜一筹。

文章首先定义了一个基于提示（Prompt Based）的分割任务，它足够普遍，能够提供强大的预训练目标，并能实现广泛的下游应用。这个任务需要一个支持多种提示并可以根据提示实时输出合理分割掩码的模型，以便进行交互使用。同时，为了训练这样一个模型，同样需要一个多样化、大规模的训练数据。图 1展示了论文中三个相互关联的组件。

首先，任务的设计灵感来自于自然语言处理（NLP）领域，例如 NLP 中可以通过预测下一个词语（Next-token）作为预训练任务，在下游任务中使用提示工程（Prompt Engineering）做相关应用。因此，为了建立分割基础模型，任务的设计目标也需要具有类似的能力。这篇文章的工作支持的提示类型有：点（Point），边界盒（Bounding Box），二值掩码（Mask），自由文本（Free-text）。将上面提到的多种提示提供给模型，训练目的是让模型根据提示输出对应的分割结果。

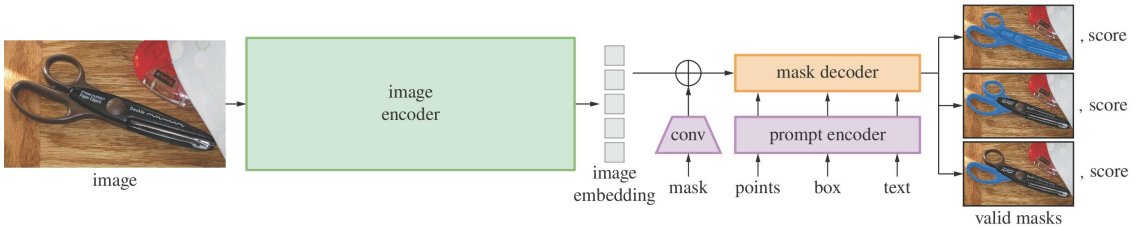


图 2. SAM 分割模型概览。重量级图像编码器输出的图像嵌入可以通过各种输入的提示进行高效查询，以实时的速度生成对象掩码。对于不明确的提示，SAM 可以输出多个有效掩码和与之关联的置信度分数。

分割模型的架构被详细地展示在图 2中。给定一图像，该图像首先被送入一个图像编码器，该编码器负责计算并生成图像的嵌入表示。与此同时，相关的提示通过一个专门的提示编码器进行处理，以产生相应的提示嵌入（Prompt Embedding）。这两种嵌入随后被送入一个设计精巧的轻量级掩码解码器，该解码器负责融合图像和提示的特征，并最终输出分割掩码。

对于给定的提示，模型被设计为输出三种不同粒度的掩码，这些掩码分别对应于目标物体的整体、部分和更细小部分的分割。这种多尺度输出能够捕捉到物体的不同层次结构，从而提供更为丰富的分割信息。输出的掩码随后会根据它们与真实标注的交并比（IoU, Intersection over Union）进行排序，这一步骤确保了模型输出的质量，使得最终的分割结果能够以一种量化的方式反映模型对于物体不同部分识别的准确性。这种方法体现了模型在处理复杂视觉和语言任务时的精细度和灵活性。

为了得到足量数据用于训练这一基础模型，论文还设计了一个数据引擎，利用数据引擎，本模型中大量的训练数据中有超过 99% 是由模型配合数据引擎自主生成。数据引擎主要分为以下三部分：

1. 辅助标注，利用可以获取到的开源数据集训练一个初始的 SAM 模型 v0 版本，再用这一模型在没有标注的数据上生成预标注，最后人工检查模型的结果并做修改和确认。

2. 半自动化标注，通过上一阶段，已经有一个不错的 SAM 模型对图像进行分割，半自动化标注是为了增加掩码的多样性。主要做法是训练一个检测模型，用于检测 SAM 生成的分割掩码是否可信，只保留可信的结果，最后把结果进行人工标注。

3. 自动标注，经历前两个阶段，SAM 可以产生不错的图像分割结果，模型此时可以利用一些筛选机制对图像进行自动标注。

3.2 论文主要创新点

本文提出了一系列技术创新，旨在推进图像分割领域的研究边界。具体技术创新点如下：

- 借鉴自自然语言处理（NLP）领域的相关研究，本文首次提出了“基于提示的图像分割”这一概念。该技术通过解析给定的 prompt 来指导图像的精确分割，使得模型能够在预训练阶段就针对广泛的下游任务进行有效的学习和适应。

- 本文设计了一个新颖的图像分割模型，该模型能够灵活地处理各种提示。模型结构包括一个图像编码器和一个多模态提示编码器，这两者的输出在一个轻量级的掩码解码器中融合，以预测最终的分割掩码。

- 提出了一种创新的“数据引擎”策略，该策略通过三个逐步演进的阶段——从 SAM 辅助的注释开始，经过部分自动化阶段，最终实现完全自动化——来收集训练数据，确保了数据在质量和多样性方面的优异性能。
- 构建并维护了 SA-1B 数据集，这是一个规模宏大的数据集，包含来自 1100 万张图片的超过 11 亿个分割掩码。这些掩码覆盖了广泛的场景，经验证具有高度的质量和多样性。SA-1B 数据集不仅为 SAM 模型的训练提供了丰富的数据资源，也为未来图像分割基础模型的研究提供了宝贵的资产。

4 复现细节

本文提出使用 SAM 进行高效的图像分割，并进一步结合 CLIP [16] 或其他多模态模型对 SAM 的输出进行进一步分析，以判断分割出的区域是否为特定的玻璃表面。同时使用本文搭建的模型在 MVS 城市大场景三维重建的图像数据上进行了多组实验和一系列分析。具体本文的复现内容如下：

- 集成 SAM 和 CLIP，利用 SAM 的分割能力和 CLIP 在文字和图像模态上的对齐能力筛选出图像中的特定掩码区域。
- 使用 MVS 场景图像进行特殊表面的检测和分割测试。

4.1 与已有开源代码对比

本文集成了 SAM 和 CLIP 的工作，两工作均已经开源，因此本文直接使用开源代码，将这两篇工作整合到一个框架之中，然后为达到处理特殊表面的检测和分割目的，本文实现了数据预处理和后处理部分的代码。

4.2 实验环境搭建

操作系统	Windows 11
集成环境	PyCharm
CPU	12thGenIntel(R)Core(TM)i9-12900K 128GB 3.20GHz 128G
GPU	RTX 3090ti 24GB
编程语言	Python
深度学习框架	Pytorch

表 1. 实验环境

4.3 创新点

特殊表面的复杂光学特性在一定程度上影响了 MVS 三维重建的模型质量。有效地检测并分割 MVS 图像中的这些特殊表面可能会提升重建质量。考虑到现有的预训练网络大多基

于特定的有监督数据集，它们可能并不适用于 MVS 的复杂城市场景。另一方面，尽管自行构建和标注数据集可以提供更为准确的检测结果，但这需要巨大的时间和经济投入。

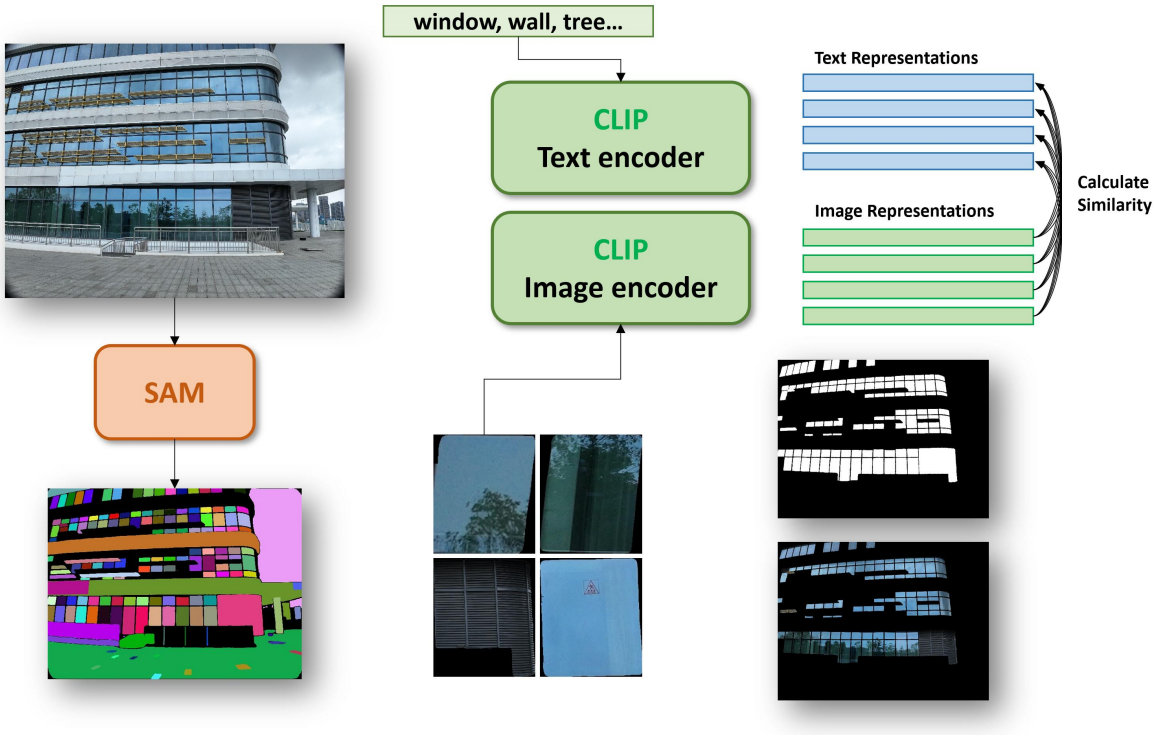


图 3. 基本流程。首先将待处理图片输入 SAM 模型进行分割处理，然后得到每一个实例的掩码。将每一个掩码输入到 CLIP 的图像编码器中，得到对应的图像表征向量。然后将预先定义好的文本输入到 CLIP 的文本编码其中得到对应的文本表征向量。最后计算每个文本编码和每个图像编码之间的相似度，最终可以得到高度语义相似性的实例掩码。

鉴于此，本文提出使用 SAM 进行高效的图像分割，并进一步结合 CLIP 或其他多模态模型对 SAM 的输出进行进一步分析，以判断分割出的区域是否为特定的玻璃表面，总体流程如图 3所示。考虑到这些大型模型在广泛场景下都展现出了卓越的泛化性能，该策略有望优化 MVS 重建中多视角对齐工作从而大幅提高城市场景中的 MVS 重建质量。

5 实验结果分析

5.1 数据准备

首先，准备具有代表性的城市大场景 MVS 重建的图片数据，如图 4，后续将使用上述图片进行实验。

图 4. 城市大场景 MVS 重建图片数据

5.2 SAM 分割

将图像输入到 SAM 模型中可以得到对图像中的每一个实例进行分割的结果，如图 5，图中的每一个色块表示 SAM 分割得到的每一个实例。

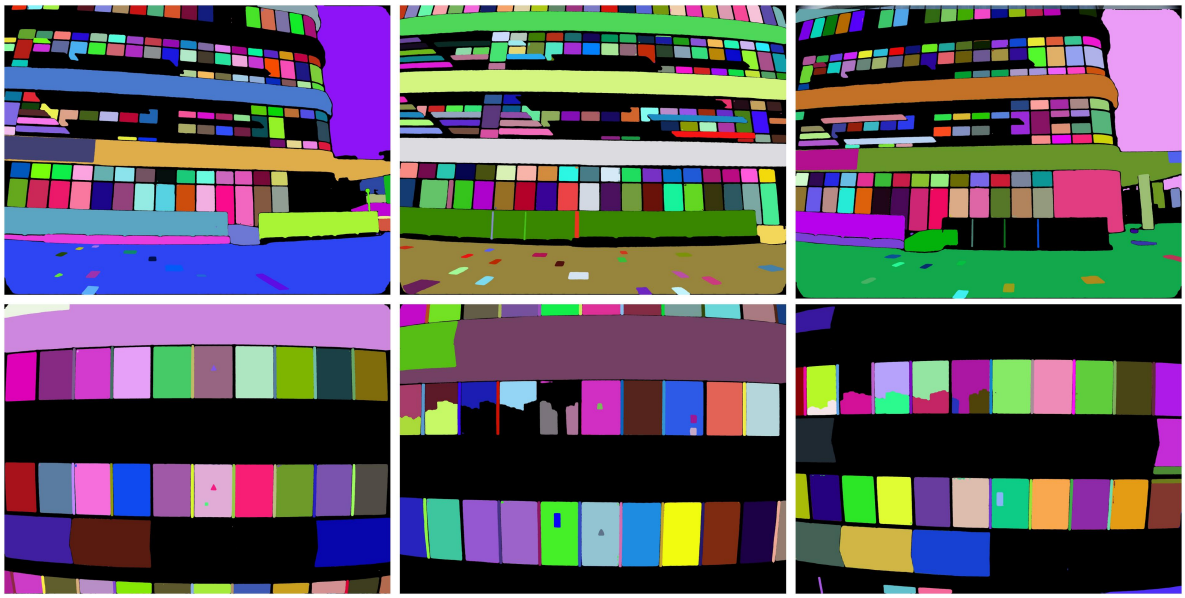


图 5. SAM 分割图

5.3 CLIP 编码

在得到输入图像的每一个实例的掩码之后，将每个实例掩码的有效区域单独切割出来，为了保留更多环境信息以便 CLIP 进行推理，在进行切割时往外扩展一定的像素值（本文往外扩展像素数量为 100）。之后将切割得到的图像输入到 CLIP 图像编码器中进行编码，并对每一个实例的编码进行记录。

同时，选取一定的语义编码输入到 CLIP 的文字编码器中，经过实验，本文挑选的一组表现较好的语义为：Window, Glass, Plant, Wall, Tree, Road, Car。

得到语义编码和图像的编码之后，直接对编码求解相似度，本文采用的是点乘相速度，每一个实例与每一个语义之间都会有一个相似度分数。最后若实例与 Window 或 Glass 相似度最高时，被认为是玻璃表面。

5.4 结果展示

通过本文搭建的模型，我们对一系列大场景照片进行了处理，如图 **??**所展示，最左侧是原始输入的图像，中间是使用 SAM 分割之后得到的实例分割图，最右侧是使用 CLIP 模型进行语义归纳之后得到的最终结果。

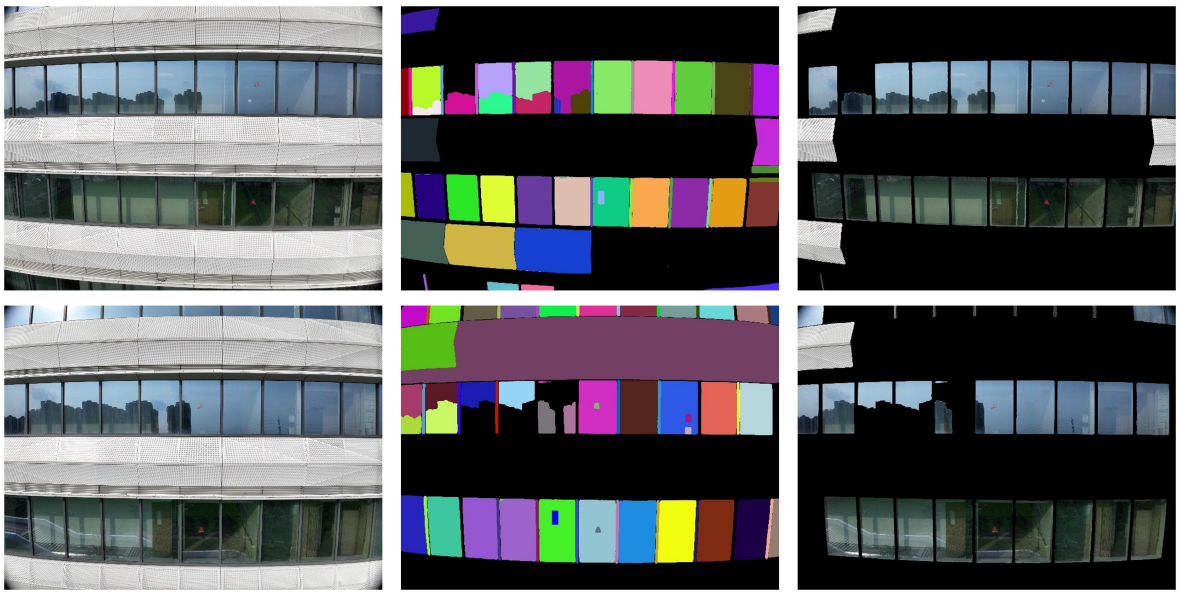


图 6. 场景 1：结果展示

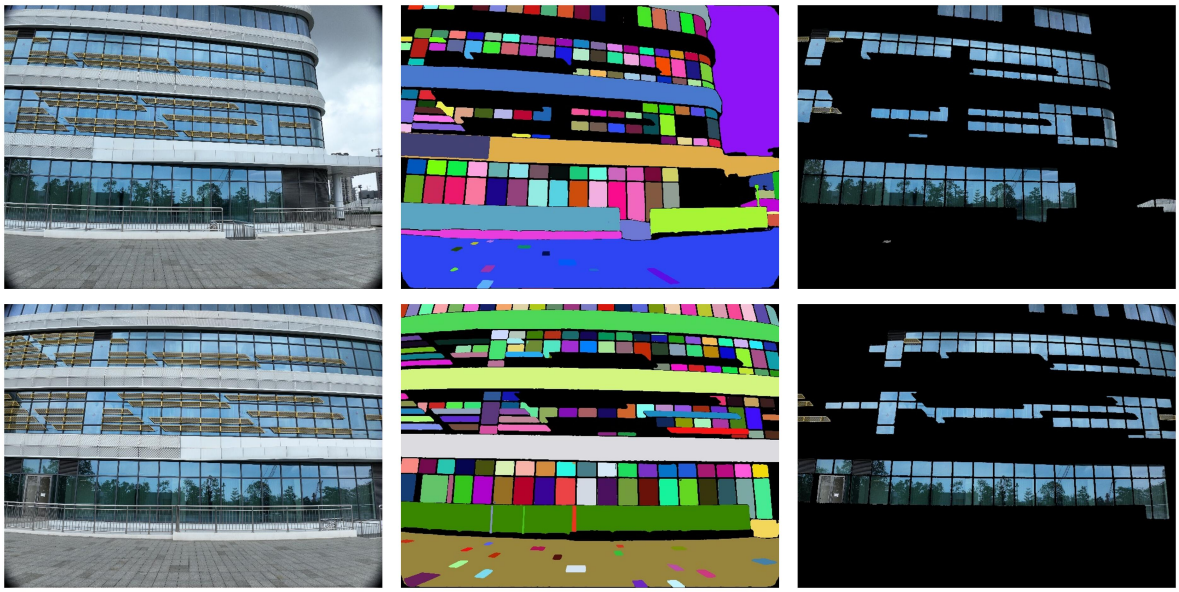


图 7. 场景 2：结果展示

通过展示的结果可以看出，本文的方法在一定程度上能够将复杂图片中的玻璃表面很好的检测和分割，这在一定程度上避免了需要在有监督情况下训练玻璃检测网络的需求。

6 总结与展望

在多视角立体视觉重建（MVS）领域，场景内部的特殊表面材质对重建精度构成了影响。针对这些表面进行精确识别和分割是降低其对三维重建精度影响的有效策略。传统的玻璃或镜面识别方法大多采取有监督学习策略，即训练特定的深度神经网络以处理图像数据。然而，鉴于现有数据集主要涵盖小范围场景的图像，这些预训练模型往往无法无缝迁移到大规模场景图像处理任务上。

本研究提出采用分割一切模型（SAM）来执行高效的图像实例分割，并辅以多模态模型 CLIP 进行语义归纳，以此准确识别图像中的玻璃区域。实验结果证实了此方法的有效性：SAM 模型能够精确分割出场景中的玻璃表面；而 CLIP 模型则能对这些分割实例进行精确的语义分析。然而，在处理特别复杂的图像场景时，SAM 模型在分辨率上的细粒度仍然存在不足，未来研究将进一步探索如何提高 SAM 在处理大规模高分辨率图像时的分割性能。

此外，通过这种新颖的检测方法识别出图像中的特殊表面后，研究将转向探索在 MVS 重建中多视角图像的特殊区域对齐。这种块对块的匹配策略预期将减少传统的点对点密集匹配所面临的挑战，从而在重建过程中显著降低由于特殊表面材质特性而产生的误差。

参考文献

[1] Haonan Dong and Jian Yao. Patchmvsnet: Patch-wise unsupervised multi-view stereo for weakly-textured surface reconstruction, 2022.

[2] Ke Fan, Changan Wang, Yabiao Wang, Chengjie Wang, Ran Yi, and Lizhuang Ma. Rfenet: Towards reciprocal feature evolution for glass segmentation, 2023.

[3] Michela Farenzena, Andrea Fusiello, and Riccardo Gherardi. Structure-and-motion pipeline on a hierarchical cluster tree. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1489–1496, 2009.

[4] Yasutaka Furukawa and Carlos Hernández. 2015.

[5] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5941–5950, June 2022.

[6] Dongshen Han and Seungkyu Lee. Internal-external boundary attention fusion for glass surface segmentation, 2023.

[7] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity, 2017.

[8] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and Lubin Weng. Enhanced boundary learning for glass-like object segmentation, 2021.

[9] Benjamin Hepp, Matthias Nießner, and Otmar Hilliges. Plan3d: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction, 2018.

[10] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson W. H. Lau, and Wangmeng Zuo. Symmetry-aware transformer-based mirror detection, 2022.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[12] Jiaying Lin and Rynson W.H. Lau. Self-supervised pre-training for mirror detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12227–12236, October 2023.

[13] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proc. CVPR*, 2020.

[14] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[15] Haiyang Mei, Xin Yang, Letian Yu, Qiang Zhang, Xiaopeng Wei, and Rynson W. H. Lau. Large-field contextual feature learning for glass detection. pages 1–17.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[17] Andrea Romanoni and Matteo Matteucci. Tapa-mvs: Textureless-aware patchmatch multi-view stereo, 2019.

[18] Shuhan Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE Transactions on Image Processing*, 22(5):1901–1914, 2013.

[19] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo, 2020.

[20] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision - 3DV 2013*, pages 127–134, 2013.

[21] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild, 2020.

[22] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer, 2021.

[23] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo, 2018.

[25] C. Zheng, D. Shi, X. Yan, D. Liang, M. wei, X. Yang, Y. Guo, and H. Xie. Glassnet: Label decoupling-based three-stream neural network for robust image glass detection, 2022.

[26] Chengyu Zheng, Peng Li, Xiao-Ping Zhang, Xuequan Lu, and Mingqiang Wei. Don't worry about mistakes! glass segmentation network via mistake correction, 2023.