Building Built in Minutes - Structure From Motion

ENPM673 Final Project

Abhishek Avhad
University of Maryland College Park
abhi6@umd.edu

Ishan Kharat
University of Maryland College Park
ishanmk@umd.edu

Piyush Goenka
University of Maryland College Park
pgoenka@umd.edu

***Abstract*—The effectiveness of estimating depth with one camera depends heavily on having extensive and varied training data. Obtaining accurate depth information across various environments at a large scale poses significant challenges, resulting in the emergence of several datasets with unique characteristics and biases. Our solution involves developing tools that allow for the integration of multiple datasets during training, even when their annotations are not directly compatible. Specifically, we introduce a robust training objective that remains consistent despite variations in depth range and scale, advocate for principled multi-objective learning to merge data from different sources, and emphasize the importance of pretraining encoders on supplementary tasks. Utilizing these tools, we conduct experiments with five diverse training datasets, including a novel, extensive dataset derived from 3D films. To showcase the versatility of our approach, we perform zero-shot cross-dataset transfer evaluations on datasets not encountered during training. The results affirm that blending data from complementary sources significantly enhances monocular depth estimation. Our method distinctly outperforms alternative techniques across various datasets, establishing a new state-of-the-art standard for monocular depth estimation..**

***Index Terms*—— Monocular depth estimation, Single-image depth prediction, Structure From Motion**

I. Introduction

The field of Structure-from-Motion (SfM), which involves reconstructing three-dimensional scenes from two-dimensional images, has undergone significant advancements, particularly in handling unordered image collections. Early systems focused on self-calibrating metric reconstruction, laying the groundwork for subsequent applications in diverse contexts such as Internet photo collections and urban environments. Inspired by these foundational works, researchers have developed increasingly sophisticated reconstruction systems capable of handling datasets ranging from hundreds of thousands to millions and even up to a hundred million images from the internet.

A variety of strategies have emerged in the realm of SfM, including incremental, hierarchical, and global approaches. Incremental SfM, characterized by its step-by-step reconstruction process, has become particularly popular for handling unordered photo collections due to its practicality and efficiency.

However, despite the significant progress made in SfM techniques, several challenges persist. These include ensuring robustness, accuracy, completeness, and scalability in the reconstruction process. These challenges hinder the widespread adoption of incremental SfM as a general-purpose method for diverse applications.

This report investigates three different methodologies for Structure from Motion (SfM), a pivotal technique in reconstructing three-dimensional scenes from a sequence of two-dimensional images. We delve into each approach, starting with Classic OpenCV Techniques, which relies on established computer vision algorithms within the OpenCV library. This method employs feature detection, matching, and triangulation to estimate the 3D structure.

Moving forward, we scrutinize MiDaS (Mixed Depth and Semantics) Large Model, a state-of-the-art deep learning model designed specifically for monocular depth estimation. Trained on a diverse dataset, it excels in predicting depth maps from individual images, offering heightened accuracy and robustness compared to traditional techniques.

Lastly, we analyze COLMAP, a versatile SfM pipeline renowned for its flexibility and robustness. This approach encompasses multiple stages, including feature extraction, matching, geometric verification, and bundle adjustment, resulting in high-quality 3D reconstructions even in challenging scenarios.

Throughout this report, we meticulously examine the intricacies of each method, highlighting their respective strengths and limitations. Additionally, we present a comprehensive comparison of their performance across a range of datasets, aiming to provide valuable insights into the effectiveness of these SfM approaches.

A. Structure From Motion (SFM)

Structure from Motion (SfM) involves the reconstruction of three-dimensional structure from its projections in a series of images captured from various viewpoints. In this paper, Incremental SfM (referred to as SfM) is discussed as a sequential processing pipeline with an iterative reconstruction element. Typically, it begins with feature extraction and matching, followed by geometric verification. The resultant scene graph forms the basis for the reconstruction stage, initializing the model with a meticulously chosen two-view reconstruction. Subsequently, new images are incrementally registered, scene points are triangulated, outliers are filtered, and the reconstruction is refined using bundle adjustment (BA).

Identify applicable funding agency here. If none, delete this.

The ensuing sections provide a detailed explanation of this process, establish the notation used throughout the paper, and introduce relevant prior research.

B. Need for SFM

The burgeoning demand for 3D models across diverse sectors such as gaming, virtual reality, architecture, and robotics underscores the necessity for accurate and efficient 3D reconstruction techniques. Traditional methods, including manual 3D modeling, are fraught with drawbacks such as time-consuming processes and labor-intensive tasks. Additionally, alternatives like laser scanning and depth sensors, while capable, often pose challenges due to their high cost and limited accessibility. In contrast, Structure from Motion (SfM) offers a compelling solution by reconstructing 3D models from a collection of 2D images. Notably, SfM stands out for its cost-effectiveness and accessibility, necessitating only a camera for implementation. This approach not only mitigates financial barriers but also enables the reconstruction of large-scale scenes and objects. The versatility of SfM finds application in various domains, including cultural heritage preservation and digitization, autonomous navigation and mapping for robots and drones, virtual tourism, and immersive experiences, as well as archaeological site documentation and analysis. By leveraging SfM, these applications can benefit from enhanced efficiency, affordability, and scalability, paving the way for innovative solutions in diverse fields.

II. Methodology

A. Structure From Motion Using OpenCV

Image Acquisition and Preprocessing involves the initial steps of the Structure from Motion (SfM) pipeline, starting with the reading of a series of overlapping images from a directory. These images may need to be downscaled to enhance computational efficiency, particularly if dealing with a large dataset. Feature Detection and Matching constitute the next stage, where distinctive features are identified in each image using the Scale-Invariant Feature Transform (SIFT) algorithm. Subsequently, feature matches are established between pairs of consecutive images using a Brute-Force Matcher. To enhance the robustness of the matches, outliers are filtered based on distance ratio criteria.

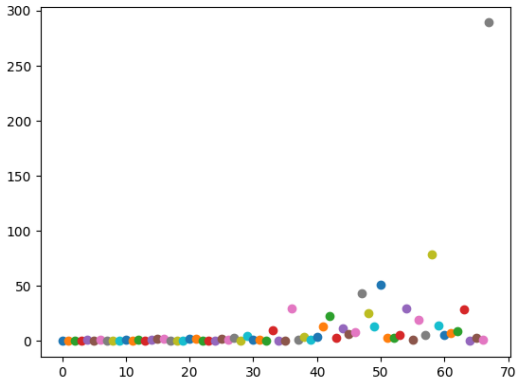

Fig. 1. Points Obtained using OpenCV

The following step involves Essential Matrix and Pose Estimation. Here, the essential matrix between the first two images is computed using the Random Sample Consensus (RANSAC) algorithm, facilitating the recovery of the relative pose (rotation and translation) between these images.

Triangulation is then employed to determine the 3D positions of the matched features using the estimated poses and camera intrinsics. Additionally, the reprojection error is calculated to evaluate the accuracy of the reconstruction.

The process transitions into Incremental SfM, where subsequent images are iterated through. Feature matches are found between the current image and the previous one, and common points are identified between these matches and the previously reconstructed 3D points. The pose of the current image is estimated using Perspective-n-Point (PnP) with these common points, facilitating the triangulation of new 3D points using pairs of current and previous images. The 3D point cloud and camera poses are then updated accordingly.

Finally, Point Cloud Registration and Visualization involves registering the 3D points and their corresponding colors from the images. The resulting point cloud is saved in PLY format, allowing for visualization and further analysis. These detailed steps collectively constitute the SfM pipeline, enabling the reconstruction of three-dimensional scenes from a series of overlapping images.

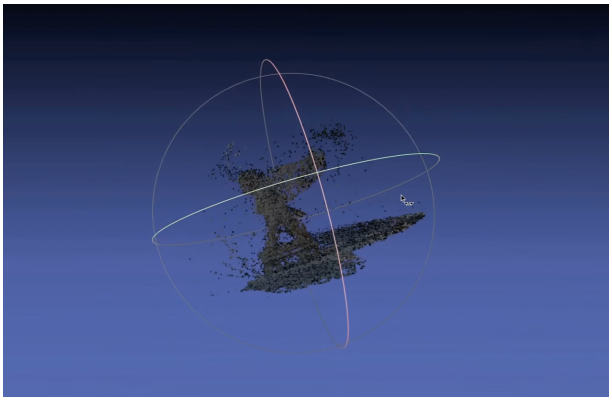

Fig. 2. Structure From Motion Using OpenCV

B. Structure From Motion Using COLMAP

Image Acquisition involves gathering a collection of overlapping images capturing the scene from various perspectives, ensuring comprehensive coverage for subsequent reconstruction.

Feature Detection and Extraction constitutes the next stage, where keypoints, or distinctive points, are identified in each image using a feature detector such as SIFT, SURF, or ORB. Alongside keypoint detection, feature descriptors are

extracted to characterize each keypoint, facilitating subsequent matching.

Feature Matching follows, where correspondences between keypoints across pairs of images are established through a nearest neighbor search. To enhance the accuracy of matches, filtering mechanisms like the ratio test or other criteria are employed to eliminate outliers. Geometric Verification is then

Fig. 3. Dataset pose using OpenCV

conducted to validate the matches obtained from the previous step. Robust estimation methods like RANSAC are utilized to discard erroneous matches and estimate fundamental or essential matrices, which encapsulate the geometric relationship between image pairs.

The process transitions into Incremental SfM, wherein the reconstruction is initialized with a single image pair. Camera poses and corresponding 3D points are estimated through triangulation. Subsequently, new images are incrementally integrated into the reconstruction framework. The pose of each new image is estimated using Perspective-n-Point (PnP) with the known 3D points, followed by triangulation of new 3D points from the matched features. Local bundle adjustment is then performed to refine the camera poses and 3D points, ensuring the coherence and accuracy of the reconstructed scene. These detailed steps collectively outline the process of Structure from Motion (SfM), enabling the creation of three-dimensional models from a series of overlapping images. Global Bundle Adjustment is a critical step in refining the en-

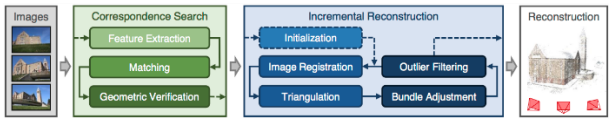

Fig. 4. COLMAP Architecture

tire reconstruction process. This optimization technique aims to minimize the reprojection error of all 3D points across all images, effectively enhancing the accuracy and coherence of the reconstructed scene. By refining both the camera poses and the positions of 3D points, global bundle adjustment ensures that the entire reconstruction aligns seamlessly with

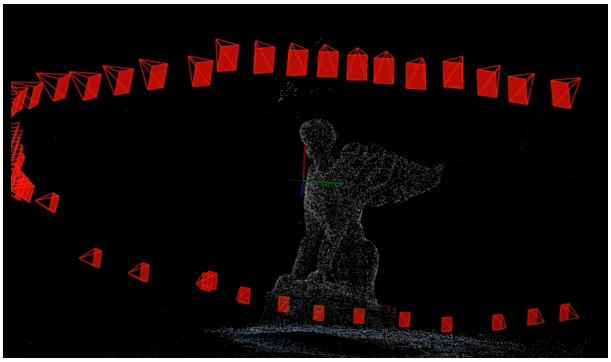

Fig. 5. Structure From Motion Using COLMAP

the observed image data. Multi-View Stereo (MVS) techniques are then employed to generate dense point clouds or depth maps from the series of images. These techniques estimate depth information for each pixel in the images, enabling the creation of a detailed representation of the scene's geometry. The depth maps are subsequently merged and refined to produce a consistent and accurate 3D model, capturing intricate details and structures.

Following the creation of dense point clouds, Mesh Generation and Refinement are carried out to construct a 3D mesh from the point cloud using surface reconstruction algorithms such as Poisson reconstruction. This process involves refining the mesh by removing outliers, filling holes, and smoothing the surface, resulting in a high-fidelity representation of the scene's geometry.

Finally, Texture Mapping is applied to add visual details to the 3D model. Texture information from the original images is mapped onto the 3D mesh, assigning color information to each vertex or face of the mesh. This step enhances the realism and visual fidelity of the reconstructed model, providing a textured representation that closely resembles the original scene.

C. Structure From Motion using MIDAS

The process begins with Image Acquisition, where a series of overlapping images of the scene are collected from different viewpoints. This ensures comprehensive coverage of the scene, providing multiple perspectives necessary for accurate 3D reconstruction. The overlapping nature of these images is crucial as it allows for the establishment of correspondences between different views, which is essential for subsequent steps.

Next, Depth Estimation is performed using the MiDaS model, a state-of-the-art deep learning model for monocular depth estimation. The pre-trained MiDaS model, which has been trained on a large and diverse dataset, is loaded into the system. Each image is then processed through the MiDaS model to generate a dense depth map, which provides an estimate of the depth value for each pixel. This transforms the 2D images into depth-augmented representations, adding a critical layer of information for 3D reconstruction.

Following depth estimation, Keypoint Detection and Description are carried out to identify distinct and repeatable

points in each image. Keypoints are detected using feature detectors such as SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features), or ORB (Oriented FAST and Rotated BRIEF). These algorithms identify salient points in the images. For each detected keypoint, a feature descriptor is extracted, capturing the local image information around the keypoints and enabling reliable matching between images.

Feature Matching is then conducted to find correspondences between keypoints in pairs of images. This involves matching the feature descriptors using nearest neighbor search techniques. To enhance the robustness of the matches, outlier matches are filtered out using techniques like the ratio test, which compares the distances of the closest and second-closest matches, or RANSAC (Random Sample Consensus), which robustly estimates geometric transformations by removing incorrect matches.

The process then moves to Depth Map Alignment, where the depth maps obtained from MiDaS are aligned with their corresponding images. Camera intrinsic parameters, such as focal length and principal point, are used to convert the depth values from the depth maps into 3D points, ensuring the depth information corresponds accurately to the image pixels. This step involves creating a 3D point cloud for each image using the aligned depth map and camera intrinsics, resulting in accurate 3D representations of the scene. Finally, Point

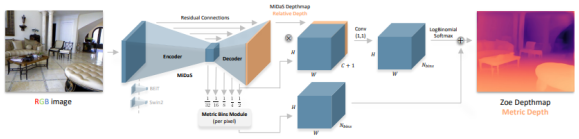

Fig. 6. MIDAS Architecture

Cloud Registration is performed to integrate the individual point clouds into a cohesive global 3D reconstruction. Using the matched keypoints and their corresponding 3D points, the point clouds are registered by estimating the relative pose (rotation and translation) between pairs of point clouds. Techniques like ICP (Iterative Closest Point) or RANSAC are employed to iteratively align the point clouds, minimizing the distance between corresponding points. Once the relative poses are estimated, the point clouds are aligned and merged, creating a consistent and accurate global 3D reconstruction of the scene.

By following these detailed steps, the process efficiently transforms a series of 2D images into a comprehensive and accurate 3D model, leveraging advanced techniques in depth estimation, feature detection, and point cloud registration. This integrated approach ensures high-quality 3D reconstructions suitable for various applications.

III. Comparative Study

A. Completeness and Complexity

OpenCV provides the fundamental building blocks for SfM, including functions for feature detection, matching, and pose estimation. However, it lacks a complete end-to-end SfM pipeline. Users must manually integrate these functions to create a coherent workflow, which can be complex and time-consuming. In contrast, COLMAP offers a comprehensive and ready-to-use SfM and Multi-View Stereo (MVS) framework. This framework includes a complete pipeline for 3D reconstruction, encompassing feature extraction, matching, incremental SfM, bundle adjustment, and dense reconstruction. COLMAP's structured and streamlined approach significantly simplifies the process, providing users with an out-of-the-box solution for high-quality 3D reconstructions.

B. Incremental SfM

While OpenCV provides the necessary functions for essential matrix estimation, pose estimation, and triangulation, users must manually implement the incremental SfM pipeline. This involves significant effort to handle the sequential addition of images, pose estimation, and triangulation, making the process cumbersome. Conversely, COLMAP features a built-in incremental SfM pipeline that automates the entire process. It efficiently manages the addition of new images, estimates camera poses, and triangulates 3D points. Additionally, COLMAP uses robust techniques for pose estimation and incorporates loop closure detection, which enhances the consistency and accuracy of the reconstruction, offering a more reliable and user-friendly experience.

C. Dense Reconstruction

OpenCV primarily focuses on sparse reconstruction and lacks built-in functions for dense reconstruction or multi-view stereo. Users must implement their own dense reconstruction pipeline using other libraries or algorithms, which can be complex and prone to errors. On the other hand, COLMAP includes a multi-view stereo module that performs dense reconstruction. This module generates dense point clouds or depth maps from the sparse SfM output, providing a seamless transition from sparse to dense reconstruction. COLMAP's MVS pipeline is highly optimized, producing high-quality dense reconstructions that capture intricate details of the scene, making it a superior choice for comprehensive 3D modeling tasks.

IV. Results

A. Comparison Between COLMAP and OpenCV

In our comparative analysis of COLMAP and OpenCV, we discovered that COLMAP is significantly more efficient and accurate. The key advantage of COLMAP lies in its ability to iteratively reduce projection error during the reconstruction process. In each iteration, COLMAP refines the estimated camera poses and 3D point positions through bundle adjustment, a process that minimizes the discrepancies between the observed image points and the reprojected points from the 3D model. This iterative refinement leads to a more accurate and coherent 3D reconstruction, making COLMAP a superior choice for tasks requiring high precision and reliability.

B. Depth Estimation Through Triangulation and MVS

The use of triangulation and Multi-View Stereo (MVS) techniques in our workflow has enabled us to achieve depth estimation effectively. Triangulation involves using multiple viewpoints to calculate the 3D positions of matched feature points, providing an initial sparse reconstruction of the scene. MVS then builds on this sparse reconstruction by generating dense depth maps or point clouds. This process captures more detailed depth information across the entire scene, enhancing the overall quality and completeness of the 3D model. By integrating these methods, we can produce robust depth estimates that contribute to high-fidelity 3D reconstructions.

C. Efficiency of MiDaS with Live Datasets

When utilizing MiDaS for depth estimation with live datasets captured from a PC camera, we achieved highly efficient results. MiDaS, known for its deep learning-based approach to monocular depth estimation, provided dense depth maps that were notably accurate and detailed. Compared to other methods we discussed, MiDaS demonstrated superior performance in terms of depth estimation and the generation of point clouds. The efficiency and effectiveness of MiDaS in processing real-time data make it an excellent tool for applications that require immediate depth and 3D information, such as live streaming or interactive environments.

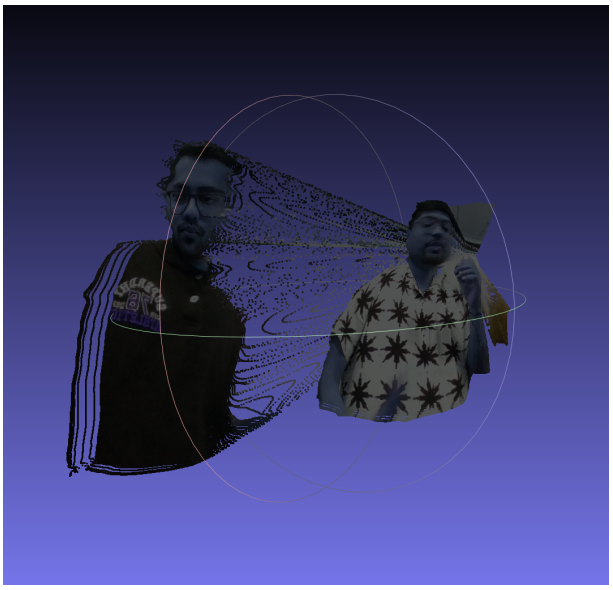

Fig. 7. MIDAS Result

D. Application in Drones and Autonomous Vehicles (AVs)

MiDaS's efficiency and accuracy make it highly suitable for gathering live data in drones and autonomous vehicles (AVs). Unlike COLMAP and OpenCV, which are typically more resource-intensive and may require more extensive computational infrastructure, MiDaS offers a streamlined approach that can be implemented in real-time scenarios. Its ability to provide rapid depth estimation and generate accurate point clouds from a single camera feed is particularly advantageous for drones and AVs, where quick and reliable data processing is crucial for navigation and decision-making. This real-time capability, combined with its robustness and accuracy, positions MiDaS as a more efficient solution compared to traditional methods like COLMAP and OpenCV for live data collection in dynamic and mobile environments.

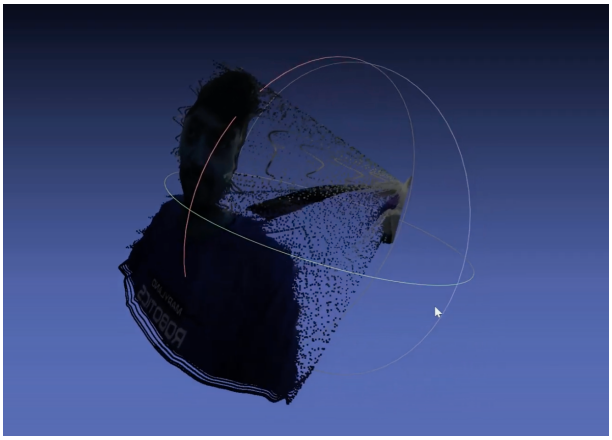

Fig. 8. MIDAS Result

V. Future Work

At present, our capabilities are limited to generating point clouds from a front perspective. This means that our 3D reconstructions are currently constrained to views from a single angle, which limits the comprehensiveness of the models we can produce. Despite this limitation, we see significant potential in the MiDaS framework for advancing our research and expanding our capabilities.

MiDaS, a state-of-the-art deep learning model for monocular depth estimation, has shown impressive results in producing dense depth maps from single images. Its ability to generate detailed depth information quickly and accurately makes it a promising tool for enhancing our 3D reconstruction processes. By leveraging the strengths of MiDaS, we aim to move beyond front perspective point clouds and work towards generating 360-degree point clouds.

The potential of MiDaS lies in its robust depth estimation capabilities, which, when further developed and integrated with advanced reconstruction techniques, could enable us to capture and reconstruct scenes from all viewpoints. This would involve not only refining the depth estimation process but also improving the methods for aligning and merging depth maps from multiple perspectives. The ultimate goal is to create a seamless and complete 3D model that accurately represents the entire scene from every angle.

As we continue our research with MiDaS, we expect to develop new algorithms and methodologies that will allow us to fully utilize its capabilities. This includes enhancing the alignment of depth maps, improving the accuracy of 3D point cloud generation, and integrating these advancements into a cohesive 360-degree reconstruction pipeline. With time and

continued effort, we are confident that MiDaS will enable us to achieve comprehensive 360-degree point clouds, significantly advancing the field of 3D reconstruction.

VI. Conclusion

Based on our analysis and comparison of different Structure from Motion (SfM) techniques, we have drawn several key conclusions about their respective strengths and limitations.

OpenCV techniques are well-suited for basic SfM tasks and serve as excellent tools for educational purposes. These traditional computer vision methods, including feature detection, matching, and triangulation, are implemented using algorithms available in the OpenCV library. While these techniques can provide a solid foundation for understanding the principles of SfM, they often require significant manual effort to implement effectively. This includes tuning parameters, handling outliers, and optimizing the workflow, which can be labor-intensive and time-consuming. Consequently, OpenCV techniques are ideal for small-scale projects and learning environments where the emphasis is on grasping the underlying concepts rather than achieving high-performance results.

The MiDaS model, on the other hand, offers a modern, automated approach to obtaining dense depth maps from single images. This deep learning model excels at producing detailed depth information quickly, making it a valuable tool for applications that require rapid depth estimation. However, while MiDaS provides an impressive initial output, the resulting depth maps often require additional refinement steps to be usable in a comprehensive SfM pipeline. This might involve aligning the depth maps with corresponding images, converting depth values to 3D points using camera intrinsics, and integrating these points into a coherent 3D model. Thus, while MiDaS significantly streamlines the depth estimation process, it still necessitates further processing to achieve the high-quality results needed for complete 3D reconstructions.

COLMAP stands out as the most comprehensive and high-quality solution for large-scale 3D reconstruction tasks. This advanced SfM pipeline includes sophisticated features and optimizations that surpass those offered by basic OpenCV techniques and automated depth estimation models like MiDaS. COLMAP's robust workflow encompasses multiple stages: feature extraction, feature matching, geometric verification, incremental and global bundle adjustment, and dense reconstruction using multi-view stereo (MVS). These capabilities allow COLMAP to handle complex and large-scale scenes effectively, producing highly accurate and detailed 3D models. The software's ability to manage challenging scenarios, such as varying lighting conditions and diverse viewpoints, further underscores its superiority. Consequently, COLMAP is particularly well-suited for professional applications and research projects that demand precise and extensive 3D reconstructions.

In summary, while OpenCV techniques are beneficial for basic tasks and educational contexts, they require considerable manual effort. MiDaS provides a quick and automated method for depth estimation but needs additional refinement for full SfM integration. COLMAP, with its advanced features and optimizations, offers the most thorough and high-quality solution for large-scale 3D reconstruction, making it the preferred choice for complex and professional applications.

Acknowledgment

This work was inspired by Johannes L. Schonberger, and Jan-Michael Frahm1's paper, "Structure-from-Motion Revisited" Special thanks to the University of Maryland, College Park and the Maryland Applied Graduate Engineering Program for their support

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building rome in a day. ICCV, 2009.

[2] S. Agarwal, N. Snavely, and S. Seitz. Fast algorithms for L problems in multiview geometry. CVPR, 2008..

[3] S. Agarwal, N. Snavely, S. Seitz, and R. Szeliski. Bundle adjustment in the large. ECCV, 2010.

[4] C. Aholt, S. Agarwal, and R. Thomas. A QCQP Approach to Triangulation. ECCV, 2012

[5] C. Beder and R. Steffen. Determining an initial image pair for fixing the scale of a 3d reconstruction from an image sequence. Pattern Recognition, 2006

[6] B. Zhou, P. Krahenb ¨ uhl, and V. Koltun, "Does computer vision matter ¨for action?" Science Robotics, vol. 4, no. 30, 2019..

[7] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," ACM Transactions on Graphics, vol. 24, no. 3, 2005