





# Image2Reality: Neural Rendering for 3D Creation

A segmentation-augmented pipeline for cleaner and faster reconstructions

**Abdullah Altamimi**
*AbdulAltamimi246@gmail.com*
*Contribution: Segmentation, preprocessing, deployment (attempted)*

**Hussain Mohsen**
*1hussain.alyafei@gmail.com*
*Contribution: Team leader, diffusion (attempted), deployment (attempted)*

**Abdulaziz Albabtain**
*abdullazizfab@gmail.com*
*Contribution: Data collection, segmentation, web*

**Abdullah Alghuzaa**
*abdllh1771@gmail.com*
*Contribution: Data collection, preprocessing, segmentation, model, diffusion (attempted)*

**Mentor: Dr. Salman Khan**
*salmank255@gmail.com*

August 27, 2025

# Abstract

We present an automated pipeline for reconstructing 3D assets from multi-view 2D images by combining Structure-from-Motion (Pixel-Perfect Structure-from-Motion (PixSfM)+COLOmap Multi-View Stereo and SfM (COLMAP)), Gaussian Splatting (GS) for efficient rendering, and foreground segmentation (SAM2) with mask propagation and monocular depth (Depth-Anything-V2 (Depth-Anything-V2)) to suppress background clutter. Our approach improves perceptual quality relative to NeRF and Instant-NGP, while achieving faster training suitable for real-world applications.

# Contents

**Abstract** **1**

**1 Introduction** **3**

**2 Problem Statement** **3**

**3 Objectives** **3**

**4 System Overview** **3**

**5 Methodology** **4**
5.1 Image Preprocessing and Segmentation . . . . . 4
5.2 Structure-from-Motion (PixSfM + COLMAP) . . . . . 4
5.3 Depth Estimation (Depth-Anything-V2) . . . . . 4
5.4 Gaussian Splatting (GS) . . . . . 4
5.5 Implementation Details . . . . . 4

**6 Evaluation** **4**
6.1 Qualitative Comparison . . . . . 4
6.2 Quantitative Metrics . . . . . 4
6.3 Discussion . . . . . 4

**7 Reflections on Challenges** **5**

**8 Related Work** **5**

**9 Conclusion** **5**

**10 Future Work** **5**

# 1 Introduction

High-quality 3D content is increasingly required for virtual showrooms, real estate visualization, and AR/VR. Neural rendering via NeRF[1] achieves high-quality view synthesis but is compute-intensive; Instant-NGP[2] accelerates training, while GS delivers real-time rendering[3]. We target common pain points by integrating interactive segmentation[4] and mask propagation into the reconstruction workflow and adopting GS for fast, photorealistic rendering.

# 2 Problem Statement

Conventional multi-view 3D reconstruction is hindered by: (i) background clutter leaking into geometry and appearance, (ii) manual cleanup steps, and (iii) long training times in neural radiance field methods[1].

# 3 Objectives

- Build an automated pipeline from image/video input to 3D output.
- Use SAM2-based interactive segmentation with mask propagation.
- Employ PixSfM+COLMAP for robust camera poses and Depth-Anything-V2 for depth cues.
- Train GS for efficient rendering of the cleaned scene.

# 4 System Overview

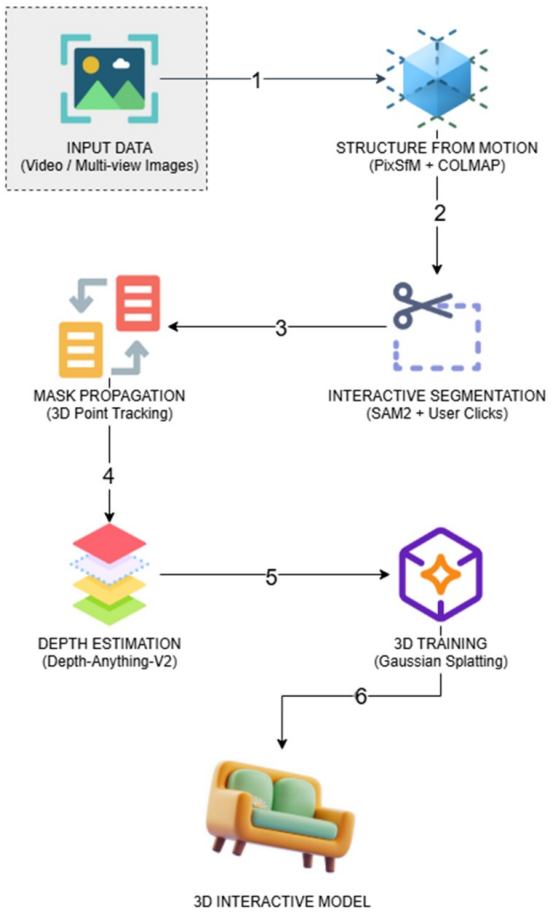



Figure 1: Updated pipeline: Input → SfM (PixSfM+COLMAP) → Segmentation (SAM2) → Mask Propagation → Depth Estimation (Depth-Anything-V2) → 3D Training (GS) → Interactive 3D Model.

The pipeline begins with input images or video, which are processed through PixSfM+COLMAP[5,6] to estimate camera poses and generate a sparse point cloud. Foreground objects are extracted using SAM2 segmentation[4], and masks are propagated across frames to reduce manual annotation. Depth maps from Depth-Anything-V2[7] add geometric cues that improve 3D consistency. Finally, the cleaned data is trained with GS[3], producing high-quality reconstructions that can be rendered interactively in real time.

# 5 Methodology

## 5.1 Image Preprocessing and Segmentation

We correct orientation/size and apply SAM2 for interactive segmentation to isolate the foreground object[4]. Masks are propagated across frames using 3D point tracking so only minimal user input is needed for a sequence.

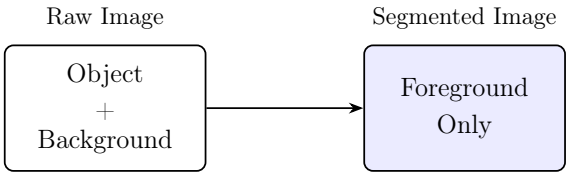



Figure 2: Illustration of segmentation workflow: raw input vs. foreground-extracted result using SAM2[4].

## 5.2 Structure-from-Motion (PixSfM + COLMAP)

We extract features, perform matching, estimate camera poses, and reconstruct a sparse point cloud followed by bundle adjustment using PixSfM+COLMAP[5,6].

## 5.3 Depth Estimation (Depth-Anything-V2)

Monocular depth predictions from Depth-Anything-V2 provide additional geometric cues and help filter residual background[7].

## 5.4 Gaussian Splatting (GS)

We initialize 3D Gaussians from SfM output and optimize appearance/opacity/scale parameters to produce a fast, high-fidelity scene[3].

## 5.5 Implementation Details

Python/PyTorch on CUDA GPUs (A100 recommended). Dependencies include SAM2[4], PixSfM[5], COLMAP[6], Depth-Anything-V2[7], and a GS implementation[3]. Input: multi-view images or video. Output: interactive 3D viewer.

# 6 Evaluation

## 6.1 Qualitative Comparison

Segmentation and mask propagation visibly reduce background leakage and clean silhouettes.



No segmentation | With segmentation + propagation

Figure 3: Qualitative comparison of reconstructions.

## 6.2 Quantitative Metrics

We evaluate using: (i) Reconstruction Time (min/scene), (ii) Success Rate (%), (iii) Mean Reprojection Error (px; lower is better).

Table 1: Quantitative summary of the reconstruction pipeline.

| Method | Time (min/scene)(GPU:A100) ↓ | Success Rate (%) ↑ | RPE (px) ↓ |
|---|---|---|---|
| With SAM2 + Propagation | ∼5 | ∼90 | ∼1.1 |

*Note: More photos increase time but generally improve accuracy and lower RPE.*

## 6.3 Discussion

Segmentation helps most when backgrounds are busy or colors overlap. Transparent/reflective objects remain challenging.

## 7 Reflections on Challenges

- Model/weights availability: occasional download failures.
- HEIC handling: conversion to PNG/JPEG required.
- Pose sensitivity: sparse viewpoints yield incomplete geometry.

## 8 Related Work

NeRF[1] achieves high-quality view synthesis; Instant-NGP[2] accelerates training. Foreground extraction via SAM[4] complements reconstruction. Our pipeline integrates segmentation/propagation with efficient GS rendering[3] while relying on mature SfM components[5,6].

## 9 Conclusion

We demonstrated a segmentation-augmented PixSfM+COLMAP+Depth-Anything-V2+GS pipeline that produces cleaner reconstructions and is practical for real assets creation.

## 10 Future Work

- Web Deployment: FastAPI backend + Streamlit/Next.js frontend with GPU inference.
- Text-Guided Segmentation: Integrate Grounding-DINO + SAM2 prompts[4].
- Large-Scale Batches: 100+ images per scene, distributed training.
- Interactive Editing: Simple masking/trim tools with re-optimization.

## References

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[2] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. (SIGGRAPH)*, 2022.

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *ICCV*, 2023.

[5] Paul-Edouard Sarlin, Viktor Larsson, Martin R. Oswald, Marc Pollefeys, et al. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021.

[6] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[7] Jianfeng Yang, Yixuan Wang, Yifan Jiang, et al. Depth anything V2. *arXiv preprint arXiv:2407.01491*, 2024.