Gen3DSR: Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View

Andreea Ardelean
andreea.ardelean@fau.de
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany

Mert Özer
mert.oezer@fau.de
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany

Bernhard Egger
bernhard.egger@fau.de
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany

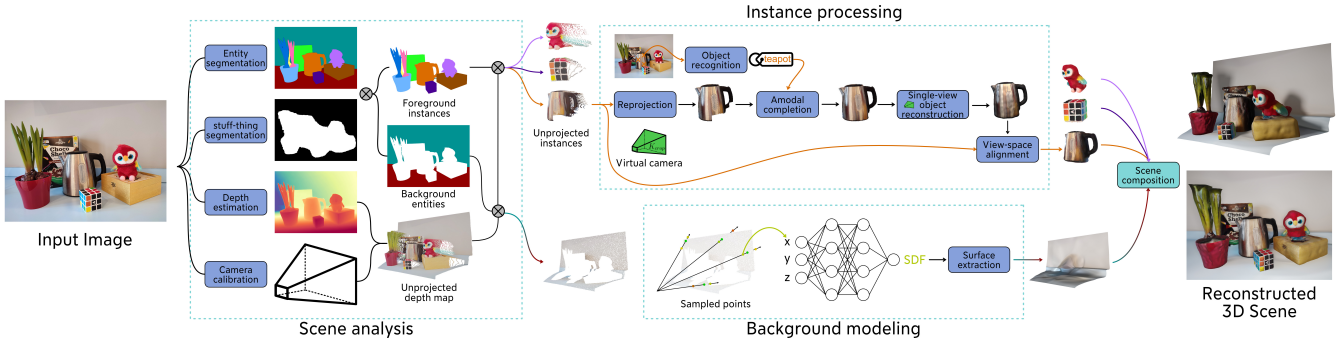

Figure 1: Method Overview: the image is first analyzed holistically using monocular models to estimate depth, semantics, and camera parameters. Scene components are then identified and categorized into foreground objects and background regions. Each foreground instance is reconstructed individually, and arranged into the scene using the unprojected depth map as a layout guide. The background is modeled separately and completes the full 3D scene composition.

ACM Reference Format:
Andreea Ardelean, Mert Özer, and Bernhard Egger. 2025. Gen3DSR: Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View. In *Proceedings of ACM Celebration of Women in Computing (womENcourage Posters '25).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The task of 3D scene reconstruction from a single view aims to understand and produce a plausible representation for the underlying geometry and appearance of all the visible components in an image. This is an ill-posed problem as there is no unique 3D solution that can explain the 2D observation. In addition, the high variety of real-world environments further increases the complexity of the task. Recently, there has been a significant progress in solving the problem for simple scenes containing isolated objects captured under predefined conditions [9, 13]. Though these methods struggle to reconstruct large scenes or objects captured under more realistic conditions (*e.g.*, object not centered in view), they achieve a good level of generalization across object categories due to their extensive training datasets [2, 3]. Following a divide and conquer strategy, we overcome the limitations of such methods, by integrating them into a modular 3D scene reconstruction framework that decouples the object-agnostic from the object-specific processing. By turning images of real-world scenes into virtual environments, the proposed method enables valuable applications in education. For example, it can be used to create interactive tools for learning physics concepts or to generate 3D assets for game development and computer graphics studies. This poster summarizes our recent work, Gen3DSR [1].

2 Method

Our method takes as input a single image and reconstructs the complete 3D scene as a collection of triangle meshes, through a three-stage pipeline illustrated in Figure 1: scene analysis, instance processing, and background modeling. Importantly, our method is not constrained to specific implementations of its modules, allowing each processing step (e.g., depth estimation, amodal completion) to be independently upgraded as improved models become available.

During scene analysis, we obtain our layout guide by unprojecting a monocular depth estimation (e.g., Metric3D v2 [6]) of the image using predicted camera calibration parameters (e.g., Perspective Fields [7]). Unlike previous works [5, 8] that rely on 2D object detectors to identify the components, we use an entity segmentation approach [10], which detects all semantically meaningful entities without being limited to a fixed set of predefined classes.

The amorphous "stuff" entities in the scene (e.g., walls and ceilings) are represented using a signed distance function (SDF) fitted to their corresponding regions in the unprojected depth map. This function is learned by a multi-layer perceptron (MLP), trained per

womENcourage Posters '25, Brașov, Romania

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX