# Generalizable 3D Foundation Models: Advancements in Geometry and View Synthesis

**Shouchen Zhou**
Tsinghua University
2025213446
zhousc25@mails.tsinghua.edu.cn

## Abstract

Recent years have witnessed a paradigm shift in 3D computer vision, moving from per-scene optimization methods to generalizable 3D foundation models. Unlike previous approaches that require training on individual scenes (e.g., NeRF[1], Gaussian Splatting[2]), these emerging models leverage large-scale pre-training and Transformer architectures to perform 3D tasks in a robust, zero-shot manner. This project proposes a comprehensive review of this development. I will analyze 4 representative papers published in 2024-2025, focusing on two key pillars: geometric reconstruction (DUSt3R, VGGT) and novel view synthesis (LVSM, RayZer). The report aims to discuss the architectural innovations that enable these models to serve as general-purpose 3D vision backbones.

## 1 Introduction and Motivation

Traditional 3D computer vision has long relied on explicit geometric solvers or per-scene optimization. While effective, these methods often struggle to generalize to unseen environments without retraining.

In contrast, the period of 2024-2025 has seen the rise of "3D Foundation Models." These models treat 3D tasks as data-driven inference problems, utilizing massive datasets to learn priors that apply across diverse scenarios.

My motivation for this project is to explore the current state-of-the-art in this domain. I aim to answer the following questions:

- How do modern foundation models implicitly learn 3D geometry from large-scale 2D image collections?
- What are the architectural advantages of using Transformers for direct 3D regression compared to traditional pipelines?
- How do recent generative approaches achieve photorealistic view synthesis with minimal 3D inductive bias?

## 2 Proposed Scope and Methodology

I plan to structure my review around representative papers that define the current state of 3D foundation models. The analysis will be divided into two main categories:

Proposal of 84761063 Deep Learning Final Project.

### 2.1 Part 1: Generalizable Geometry Reconstruction

This section will focus on models that predict 3D structure (cameras, points, depth) directly from images.

- **DUSt3R (CVPR 2024) [3]:** I will analyze how this model reframes Multi-View Stereo (MVS) as a regression task, outputting dense 3D point maps directly without prior camera information.
- **VGGT (CVPR 2025) [4]:** As an evolution of DUSt3R, I will discuss how VGGT introduces a scalable, feed-forward Transformer architecture that jointly estimates camera poses, depth, and point tracks for hundreds of images efficiently.

### 2.2 Part 2: Generalizable View Synthesis

This section will explore how foundation models tackle Novel View Synthesis (NVS) for unseen scenes.

- **LVSM (ICLR 2025) [5]:** I will examine this "Large View Synthesis Model" to understand how it achieves high-quality rendering by treating synthesis as a data-driven prediction task rather than a physics-based rendering problem.
- **RayZer (ICCV 2025) [6]:** I will review this work to highlight self-supervised learning approaches that address data scarcity in 3D training, enabling robust zero-shot generalization.

## 3 Conclusion

The final report will synthesize the methodologies of these papers, discussing their shared principles (e.g., the move towards minimal inductive bias) and their limitations. I will also provide insights into future directions, particularly the potential convergence of reconstruction and synthesis models.

## 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. *Communications of the ACM*, 65(1):99–106, 2021.

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[3] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[4] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[5] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

[6] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025.