Benchmarking Backbones for 3D Gaussian Splat Reconstruction

Ritama Sanyal

2023112027

IIIT Hyderabad

ritama.sanyal@research.iiit.ac.in

Keerthi Seela

2023102012

IIIT Hyderabad

keerthi.seela@students.iiit.ac.in

Abstract—This project systematically benchmarks multiple foundation backbones—including DINOv2, CroCo, MAS3R, and lightweight mobile-friendly models—integrated into a Splatt3R-style reconstruction framework. We evaluate each backbone across datasets such as CO3D, measuring improvements in photometric fidelity (PSNR, SSIM),runtime, and memory consumption. Our experiments reveal clear trade-offs: stronger foundation models provide superior reconstruction quality but at the cost of increased computational overhead, while lightweight backbones offer competitive performance for resource-constrained settings.

TEAM KR

Introduction to 3D Gaussian Splatting

Recent advancements in neural scene representations have led to 3D Gaussian Splatting (3D-GS), an efficient technique for real-time, high-fidelity 3D reconstruction and rendering. This method represents a scene as a set of anisotropic Gaussians, optimized from multi-view imagery, achieving competitive photometric accuracy and unprecedented rendering speed. Pipelines like Splatt3R and InstantSplat further enhance this foundation with learnable components for improved scene understanding and generalization.

A critical component in these pipelines is the visual backbone, responsible for extracting features from input images. While modern self-supervised and foundation models (DINOv2, CroCo, MAS3R) demonstrate strong generalization, their impact on Gaussian Splat reconstruction performance remains underexplored. This study addresses the gaps by integrating various backbones into a Splatt3R-style framework for a comprehensive comparative analysis.

Key Contributions

- Unified benchmarking framework for visual backbones.
- Integration and comparison of state-of-the-art foundation models and lightweight alternatives.
- Comprehensive study of reconstruction fidelity, runtime, and memory usage.
- Insights into failure cases and design guidelines for model selection.

Related Work: Evolution of 3D Reconstruction

1

3D Gaussian Splatting (3D-GS)

Represents scenes as collections of 3D Gaussian primitives for efficient real-time rendering. Extended to feed-forward settings with methods like pixelSplat and InstantSplat, which predict splats from image pairs or leverage geometric priors for faster reconstruction.

2

Foundation Models for 3D

DUST3R introduced pointmap regression for camera-free 3D reconstruction. MAS3R enhanced this with dense local feature matching for metric 3D accuracy. DINOv2 provides robust visual features, while CroCo v2 uses cross-view completion for binocular tasks.

3

Splatt3R

Extends MAS3R for pose-free, feed-forward 3D Gaussian Splatting from uncalibrated stereo pairs. It incorporates a Gaussian decoder and a novel loss masking strategy for strong performance on extrapolated viewpoints, achieving 4 FPS reconstruction.

4

Evaluation Metrics

Standard metrics include PSNR for pixel fidelity, SSIM for structural similarity, and LPIPS for perceptual judgments. Pose estimation uses translation and rotation errors, while geometric accuracy relies on depth metrics like RMSE.

Methodology: Benchmarking Pipeline

Our benchmarking pipeline systematically evaluates the influence of feature extraction architectures on Gaussian Splat reconstruction. We integrate four distinct backbones into the Splatt3R framework, ensuring a controlled comparison.

Feature Extraction

Backbones process input RGB images to produce dense or patchwise features. Transformer models use default patch sizes, while lightweight models produce lower-dimensional convolutional features.

Adapter Layers

A learned linear projection layer maps backbone-specific feature dimensions to a common embedding space, crucial for Gaussian initialization and optimization within Splatt3R.

Splatt3R Integration

Extracted feature maps guide Gaussian placement, opacity initialization, and attribute refinement. All backbones share the same downstream optimization pipeline for fair comparison.

Datasets and Metrics

We primarily use the **CO3D dataset**, a large-scale collection of multi-view object videos with calibrated camera poses. This dataset is ideal for multi-view reconstruction tasks.

Reconstruction Fidelity Metrics

- PSNR (Peak Signal-to-Noise Ratio)
- SSIM (Structural Similarity Index)
- MSE (Mean Squared Error)
- LPIPS (Learned Perceptual Image Patch Similarity)
- Reconstruction Loss

Computational Analysis Metrics

- Runtime per iteration during Gaussian optimization
- Peak GPU memory usage during training

Implementation Details

All code was implemented in PyTorch (1.11+) and experiments were conducted on NVIDIA GeForce RTX 2080 Ti. We ensured reproducibility through detailed documentation of backbone implementations, training hyperparameters, and data preprocessing.

1

Software & Environment

PyTorch (1.11+), torchvision, timm, numpy, scipy, Pillow, tqdm, matplotlib. CUDA toolkit was used for compiling CroCo RoPE CUDA kernels.

2

Backbone Implementations

MAS3R (reference) used its supplied implementation. DINOv2 and CroCo v2 were loaded from official/vended checkpoints. MobileNetV3-Small served as the lightweight baseline.

3

Data Preprocessing

Image pixels normalized with ImageNet mean/std. Images resized to canonical sizes for ViT-based encoders. Mild augmentations like random horizontal flip and small color jitter were applied.

4

Model I/O & Feature Shapes

ViT tokens are reshaped and projected convolution.

Reproducibility Measures

- Fixed random seeds for Python, NumPy, and PyTorch.
- Deterministic flags for PyTorch (e.g., `torch.backends.cudnn.deterministic=True`).
- Model checkpoints and optimizer states saved every epoch, along with full config and git commit hash.

Experimental Setup: Backbones and Protocols

We benchmark four distinct backbone families, varying in model capacity, pretraining objective, and inductive bias, all integrated into the same Splatt3R pipeline for a fair comparison.

MASi3R

Geometry-driven multi-view transformer, pretrained for correspondence and reconstruction.

DINOv2

Large-scale self-supervised semantic backbone (ViT-B), testing semantic priors for 3D reconstruction.

CroCo / CroCo v2

Cross-view completion backbone, designed to infer missing spatial structure, aligning with multi-view reconstruction.

Lightweight Model

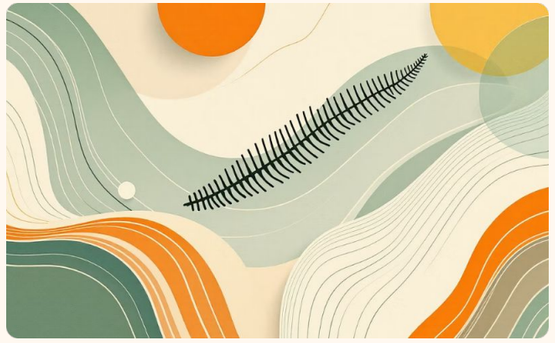
MobileNetV3-Small, an efficiency-oriented convolutional backbone for speed/memory baseline.

Training Protocol

- All models trained for 30 epochs using AdamW with cosine learning rate decay.
- Backbones frozen for the first 3–5 epochs to stabilize decoder learning.
- Encoder learning rates set lower than decoder/projection layers.
- Input images resized.
- Mixed precision used for all models to reduce GPU memory consumption.

Qualitative Results: Visualizing Reconstruction Fidelity

Visual comparisons highlight how backbone choice affects texture sharpness, structural completeness, and overall scene fidelity.



MASi3R Backbone

Produces strong baseline reconstructions with sharp edges and stable color. Minor blurring in fine details, especially with large input baselines.

DINOv2 Backbone

Captures high-level semantic structure but results in smoother reconstructions, losing high-frequency details. Optimized for semantic representation, not geometric consistency.

CroCo Backbone

Demonstrates the most visually consistent reconstructions, retaining sharper textures and fewer artifacts in occluded or sparsely observed regions, due to cross-view completion pretraining.

Lightweight Backbone

Provides fast but less detailed reconstructions. Surfaces appear smoother, and geometric boundaries are less distinct, illustrating the trade-off between efficiency and fidelity.

Figure 1: Qualitative comparison of reconstructions from different backbone architectures. CroCo and MAS3R generally preserve finer details and textures, while DINOv2 produces smoother surfaces, and the lightweight backbone sacrifices sharpness for efficiency.

Quantitative Results: Performance Metrics

We report quantitative evaluation of reconstruction fidelity using PSNR and SSIM on the held-out CO3D test split. Runtime is measured as the average time required to render a single novel view.

MAS3R (ViT-L)	27.40	0.880	0.55
DINOv2 (ViT-B)	24.60	0.830	0.42
CroCo v2 (ViT-B BaseDec)	25.70	0.850	0.38
Lightweight (MobileNetV3)	23.10	0.810	0.18

Table I: Reconstruction performance across different backbones on CO3D. Reported PSNR / SSIM are averages on the held-out category split; Runtime is average rendering time per novel view measured on an NVIDIA GeForce GTX 1080 Ti.

MAS3R achieves the highest reconstruction fidelity, while CroCo offers competitive performance with improved handling of sparsely observed regions. DINOv2 maintains strong semantic consistency but yields smoother textures, resulting in slightly lower PSNR. The lightweight MobileNetV3 backbone offers significantly reduced runtime at the cost of fine-detail accuracy.

Discussion: Trade-offs and Failure Modes

Our experiments reveal clear trade-offs between reconstruction fidelity, semantic robustness, and computational efficiency. No single backbone dominates across all dimensions; the appropriate choice depends on deployment constraints.

MAS3R (ViT-L)

Highest overall performance (PSNR, SSIM) due to large capacity and geometry-aware pretraining. However, it incurs the highest computational cost (rendering time, GPU memory).

CroCo (ViT-B BaseDec)

Strong middle ground, producing sharper textures and more stable geometry than DINOv2, especially in sparsely observed regions. Lower inference runtime than MAS3R.

DINOv2 (ViT-B)

Competitive performance but generates smoother textures and slightly lower PSNR, emphasizing semantic consistency over photometric detail. Moderate runtime.

Lightweight (MobileNetV3)

Fastest inference by a substantial margin, ideal for resource-constrained hardware. Reduced representational capacity leads to lower fidelity for fine structures and textures.

Common Failure Modes

Extreme Viewpoint Sparsity: All backbones exhibit surface thinning, blurred texture boundaries, or color bleeding, with severity correlating to model capacity.

These observations highlight the importance of aligning backbone selection with application-specific requirements, particularly for large-scale 3D reconstruction or on-device inference.

Conclusion and Future Work

Our study demonstrates that backbone choice significantly impacts 3D Gaussian Splatting performance, affecting texture fidelity, structural completeness, and robustness to sparse observations. Reconstruction-oriented pretraining (MAS3R, CroCo) leads to more faithful results compared to purely semantic or efficiency-focused backbones.

01

Extended Datasets

Evaluate on larger, more diverse multi-view datasets (e.g., BlendedMVS, DTU) to assess generalization and robustness.

02

End-to-End Training

Jointly optimize backbone and splatting renderer for improved performance, especially for backbones not originally designed for geometric tasks.

03

Geometric Metrics

Incorporate quantitative geometric evaluations (Chamfer Distance, F-score, reprojection error) for comprehensive accuracy assessment.

04

Lightweight Deployment

Design custom compact backbones or distill larger models for real-time 3D reconstruction on edge devices.

05

Temporal Consistency

Extend pipeline to video sequences with explicit temporal modeling to reduce flickering and enhance stability.

06

Hybrid Feature Fusion

Combine semantic-rich features (DINOv2) with geometry-aware features (CroCo, MAS3R) for balanced representations.

We would like to express sincere gratitude to Prof. Madhav Krishna for his guidance, insightful discussions, and continuous support throughout the course of this project. We also thank the teaching assistants Samyak Mishra, Soham Patil, and Ajit Srikanth for their valuable feedback, clarifications, and assistance during the implementation and experimentation phases. Their collective support greatly contributed to the successful completion of this work.