Scene-Level Reconstruction with Sparse Inputs

Post-Ph.D. Research Proposal

Chen Yang

Advisor: Prof. Wei Shen

Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

November 2024

A Proposal for the Future Research Directions of Chen Yang

December 10, 2024

Contents

1 Introduction

3D scene reconstruction from images has emerged as a fundamental technology that underpins numerous real-world applications. In autonomous driving, accurate 3D scene understanding enables vehicles to navigate safely and make informed decisions. In augmented reality (AR), reconstructed 3D environments allow virtual objects to interact naturally with the physical world. For digital content creation, efficient 3D reconstruction tools can dramatically reduce the time and cost of creating virtual assets for games, movies, and the metaverse.

Traditional 3D reconstruction approaches typically require hundreds of densely captured images to achieve high-quality results. However, in many practical scenarios, obtaining such dense captures is either impossible or highly impractical. For instance, autonomous vehicles must make decisions based on sparse temporal observations, AR applications expect immediate reconstruction from just a few casual photos, and historical preservation projects often only have access to a limited number of archive photos. Recent advances in neural rendering, particularly Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS), have revolutionized the field by enabling photorealistic novel view synthesis. These methods achieve impressive visual quality through implicit or explicit neural representations and differentiable rendering. However, they require extensive per-scene optimization, typically taking hours for NeRF and over 10 minutes for 3DGS. Moreover, they rely heavily on accurate camera poses from structure-from-motion pre-processing and dense image captures, making them impractical for many real-world applications.

Sparse-view reconstruction faces several fundamental challenges that make it particularly difficult. ***1) Geometric ambiguity arises from insufficient parallax and viewpoint coverage.*** With only a few input views, multiple 3D geometries could potentially explain the same observations, especially in textureless regions or areas viewed from limited angles. ***2) The sparsity of views often means large baselines between cameras***, leading to significant appearance variations and occlusions that complicate feature matching and correspondence establishment. ***3) The reconstructed geometry tends to be incomplete due to unobserved regions***, requiring effective mechanisms to reason about and fill in missing scene content.

Prior works have explored various approaches to inject additional priors (*e.g.* total variation, depth, entropy, semantics) into sparse-view reconstruction to constrain the solution space. These approaches have demonstrated impressive results for per-scene optimization in controlled settings. However, these prior-based methods face significant limitations for practical applications. The underlying priors are often not universally applicable across diverse real-world scenarios. Semantic priors may fail when encountering novel object categories or unusual scene compositions. Geometric regularization terms that work well for man-made environments might not generalize to natural scenes. Moreover, the reliance on per-scene optimization incurs substantial computational costs, typically requiring minutes or hours to process each new scene. This combination of limited generalizability and slow processing speed hinders their adoption in applications requiring robust and efficient reconstruction.

In this proposal, I choose to leverage 3D pre-trained models to address the aforementioned challenges. Pre-trained models have demonstrated remarkable capabilities in learning rich geometric and semantic priors from large-scale datasets, offering several key advantages for sparse-view reconstruction. First, they can encode comprehensive scene understanding that generalizes across diverse environments, potentially resolving geometric ambiguities through learned structural priors. Second, their feed-forward nature enables efficient inference without expensive per-scene optimization, reducing reconstruction time from hours to seconds. Third, pre-trained models can integrate multi-modal information, from geometric relationships to semantic understanding, providing a more holistic approach to scene reconstruction. Building upon these advantages, there are currently three different ways to achieve the target, and I will demonstrate their advantages and disadvantages in the **Proposed Research** section.

2 Related Work

Sparse View Reconstruction with Geometric Priors Vanilla NeRF (Mildenhall et al., 2020) typically require dense view sampling for high-quality reconstruction. To address sparse view scenarios, various geometric priors have been explored. Several methods leverage Structure from Motion (SfM) (Schönberger and Frahm, 2016) derived information, such as visibility maps or depth estimates (Deng et al., 2022; Roessle et al., 2022; Somraj and Soundararajan, 2023; Somraj et al., 2023, 2024). While effective, these approaches primarily work with closely aligned views. Alternative approaches utilize depth information, either from ground truth depth maps (Xu et al., 2022) or monocular depth estimation models (Song et al., 2023a; Guangcong et al., 2023; Ranftl et al., 2022, 2021), though the latter often produce results too coarse for detailed reconstruction.

Learning-based Priors and Regularization Various learning-based priors have been proposed to improve sparse view reconstruction. Shi et al. (2024a) combines deep image priors with factorized NeRF to capture overall appearance, though fine details may be lost. Jain et al. (2021) leverages vision-language models (Radford et al., 2021) for novel view synthesis, but the semantic guidance proves too abstract for accurate low-level reconstruction. Other approaches explore priors based on information theory (Kim et al., 2022), continuity (Niemeyer et al., 2022), symmetry (Seo et al., 2023), and frequency regularization (Yang et al., 2023; Song et al., 2023b), though their effectiveness is often limited to specific scenarios. Recent methods have also incorporated Vision Transformers (ViT) (Dosovitskiy et al., 2021) to reduce the requirements for NeRFs and Gaussians (Jiang et al., 2024; Jang and Agapito, 2024; Xu et al., 2024a; Zou et al., 2024).

Diffusion Models for 3D Reconstruction The emergence of diffusion models has revolutionized 3D reconstruction through their powerful generative capabilities. Dreamfusion (Poole et al., 2023) pioneered this direction by introducing Score Distillation Sampling (SDS), which enabled the distillation of 2D diffusion priors into NeRFs for text-to-3D generation. This breakthrough sparked numerous extensions in text-to-3D synthesis (Lin et al., 2023; Metzer et al., 2023; Wang et al., 2023a; Chen et al., 2023; Wang et al., 2023b; Yi et al., 2024; Tang et al., 2024b; Shi et al., 2024b) and 3D/4D editing applications (Haque et al., 2023; Shao et al., 2024).

For sparse view reconstruction, several approaches have been developed. Single-image methods (Zhu and Zhuang, 2024; Chan et al., 2023; Liu et al., 2023a; Burgess et al., 2024; Pan et al., 2024; Müller et al., 2024) leverage diffusion models to hallucinate novel views, though they often struggle with input constraints and image saturation. To address occluded regions, GaussianObject (Yang et al., 2024) introduces a fine-tuned Control-Net (Zhang et al., 2023) as a repair model. Other methods (Wynn and Turmukhambetov, 2023; Zhou and Tulsiani, 2023; Liu et al., 2023b; Wu et al., 2024) explore fine-tuning text-to-image models for simultaneous multi-view generation, effectively transferring knowledge from 2D image-space priors.

Recent advances in Latent Video Diffusion Models (LVDM) have opened new possibilities for sparse view synthesis. Cat3D (Gao et al., 2024) builds upon video and multi-view diffusion models to generate highly consistent novel views, while ViewCrafter (Yu et al., 2024) finetunes Stable Video Diffusion to condition on differentiable point rasterization for realistic view synthesis.

Feed-forward Methods Recent advances in sparse view reconstruction have led to two major directions: Large Reconstruction Models (LRMs) and neural matching approaches. LRMs (Hong et al., 2024b; Wang et al., 2024a; Xu et al., 2024c; Wei et al., 2024; Xu et al., 2024b; Li et al., 2024; Tang et al., 2024a; Weng et al., 2023; Zhang et al., 2024) offer impressive speed through direct feed-forward generation. Feed-forward reconstruction methods have shown impressive performance and inference speed among sparse and single-image reconstruction. However, they currently face several limitations: sensitivity to view distributions and object placements, challenges in handling real-world scenarios, and generally lower quality compared to methods leveraging image-space priors.

A parallel development in neural matching has emerged to address the fundamental challenge of camera pose estimation. Traditional reconstruction methods often require precise camera parameters, which typically demand dense views to obtain reliably. DUSt3R (Wang et al., 2024b) pioneered a novel approach by predicting point maps for uncalibrated stereo pairs within a unified coordinate system through implicit correspondence searching. Building upon this foundation, MASt3R (Leroy et al., 2024) enhanced the image-matching process by predicting points in metric space, achieving superior accuracy. These methods demonstrate remarkable stereo reconstruction capabilities even with minimal view overlap. The success of pixel-aligned point maps has inspired several recent methods (Ye et al., 2024; Smart et al., 2024) to integrate DUSt3R or MASt3R architectures with Gaussian heads for direct 3D Gaussian generation. While these approaches excel at two-view reconstruction, their performance deteriorates with additional views, inheriting the multi-view limitations of their underlying DUSt3R/MASt3R frameworks. This scalability challenge remains an active area of research in feed-forward reconstruction methods.

3 Proposed Research

3.1 Task Definition and Target

Given a set of N sparse views (typically $N = 3$-9) of a scene with corresponding camera poses, our goal is to reconstruct a complete and accurate 3D representation of the scene. Formally, let $\mathcal{I} = \{I_1, ..., I_N\}$ be the input RGB images where $I_i \in \mathbb{R}^{H \times W \times 3}$, and $\mathcal{P} = \{P_1, ..., P_N\}$ be their corresponding camera poses where $P_i \in SE(3)$ represents the extrinsic camera parameters. The target is to learn a mapping function f that generates a 3D scene representation $\mathcal{S}$:

$$\mathcal{S} = f(\mathcal{I}, \mathcal{P}; \theta) \tag{1}$$

where θ represents the learnable parameters of our model, and $\mathcal{S}$ can be rendered to novel views $\hat{I}$ through a differentiable rendering function R:

$$\hat{I} = R(\mathcal{S}, P_{novel}) \tag{2}$$

The reconstructed representation $\mathcal{S}$ should satisfy the following key requirements:

1. **Geometric Accuracy**: The reconstructed geometry should accurately match the observed views and maintain consistency with real-world physics and structural constraints.

2. **Completeness**: Despite limited input views, the reconstruction should plausibly complete unobserved regions of the scene while maintaining global consistency.

3. **Efficiency**: The reconstruction process should be completed within seconds rather than minutes or hours, making it practical for real-world applications.

4. **Generalizability**: The method should work effectively across diverse scene types without requiring per-scene optimization or fine-tuning.

3.2 Methodology

Based on recent advances in 3D pre-trained models, I identify three promising approaches to address this challenging task. First, I define our scene representation $\mathcal{S}$ as a set of 3D Gaussians:

$$\mathcal{S} = \{\boldsymbol{\mu}_j, \alpha_j, \mathbf{r}_j, \mathbf{s}_j, \mathbf{c}_j\}_{j=1}^{K}, \tag{3}$$

where K is the number of Gaussians, $\boldsymbol{\mu}_j \in \mathbb{R}^3$ represents the 3D position, $\alpha_j \in (0, 1)$ denotes opacity, $\mathbf{r}_j \in \mathbb{R}^4$ is rotation in quaternion form, $\mathbf{s}_j \in \mathbb{R}^3$ indicates scaling factors, and $\mathbf{c}_j \in \mathbb{R}^3$ represents RGB color (for simplification, here I use color $\mathbf{c}_j$ instead of $\mathcal{SH}$, Spherical Harmonics).

3.2.1 Transformer-based Feed-forward Model

The transformer-based approach, represented by GS-LRM (Zhang et al., 2024) and Long-LRM (Ziwen et al., 2024), offers a direct and effective solution for sparse-view reconstruction. These methods leverage the powerful context modeling capability of transformers to learn both local and global scene priors from large-scale datasets.

Architecture Design The core of this approach is a pure transformer architecture that directly maps sparse input views to the defined scene representation through:

$$\mathcal{S} = f_\theta(\mathcal{I}, \mathcal{P}) \tag{4}$$

where $\mathcal{I} = \{I_i\}_{i=1}^{N}$ and $\mathcal{P} = \{P_i\}_{i=1}^{N}$ are input images and camera poses respectively. The posed images are first tokenized through patch embedding into tokens $\mathbf{t}_i \in \mathbb{R}^d$, with camera parameters encoded using Plücker coordinates $[\mathbf{o}, \mathbf{d}] \in \mathbb{R}^6$ that combine ray origin and direction.

Key Advantages This transformer-based approach effectively addresses the challenges of sparse-view reconstruction. The self-attention mechanism enables the model to leverage both local and global context when determining 3D structure, helping resolve geometric ambiguities. The learned feature transformations can handle large baselines and appearance variations between views. Additionally, the feed-forward nature allows fast inference ($\sim$0.23s per scene), making it practical for real-world applications.

Limitations Despite its effectiveness, this approach currently has resolution constraints (typically limited to 512×904) and requires images with accurate camera poses and same intrinsics as input. The reconstruction is also limited to observed regions within the view frustum, with limited ability to hallucinate unseen areas. Additionally, like most transformer-based methods, it requires significant computational resources for training due to the quadratic attention

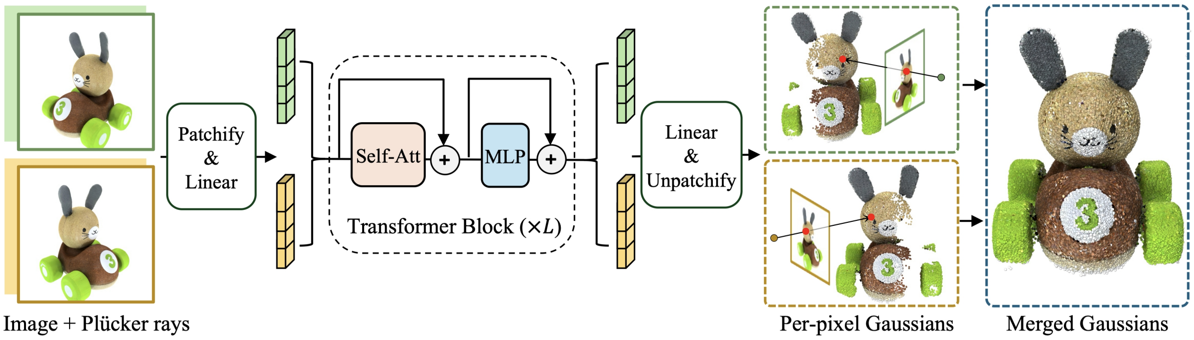

Figure 1: The pipeline of GS-LRM.

complexity. The model's performance is highly dependent on the pre-training strategy and the diversity of the training dataset. Observations on GS-LRM indicate that the synthesized images at the scene level are quite blurry and fail to accurately recover compressed information that ignored during capturing[*]. Moreover, since LRMs are proposed by Adobe and not open-sourced, current publicly available implementations cannot achieve the same level of performance as reported in their paper, making it challenging for the research community to build upon and improve this approach.

3.2.2 Multi-View/Video Diffusion Model

Video diffusion models, represented by CAT3D and ViewCrafter, provide a powerful framework for sparse-view reconstruction through consistent novel view synthesis. These models leverage the temporal coherence inherent in multi-view/video data to generate geometrically consistent novel views from sparse inputs.

Architecture Design Taking CAT3D for example, the model operates in two stages:

$$\mathbf{z}_i = E(I_i) \in \mathbb{R}^{H/8 \times W/8 \times d}, \quad I_i \in \mathcal{I} \tag{5}$$

where E is a pre-trained VAE encoder that compresses input images into a lower-dimensional latent space. The latent codes $\mathbf{z}_i$ capture high-level semantic and geometric information while reducing computational overhead.

$$\hat{\mathcal{I}} = f_\theta(\{\mathbf{z}_i, P_i\}_{i=1}^N, \{P_{novel}\}) \tag{6}$$

The diffusion model f_θ takes both the encoded latent vectors and camera poses as input to generate novel views $\hat{\mathcal{I}}$. This model employs 3D self-attention mechanisms to ensure consistency across generated views. The final scene representation $\mathcal{S}$ is reconstructed from the expanded set of views using traditional 3D reconstruction methods like NeRF or Gaussian Splatting.

Key Advantages The video diffusion approach addresses reconstruction challenges by learning consistent scene representations through joint multi-view generation. The 3D self-attention mechanism helps maintain geometric consistency across different viewpoints. It enables fast

[*] When the view is orthogonal to the surface of the object, the observed information attached to the surface can be largely preserved; On the contrary, the information will be severely compressed.

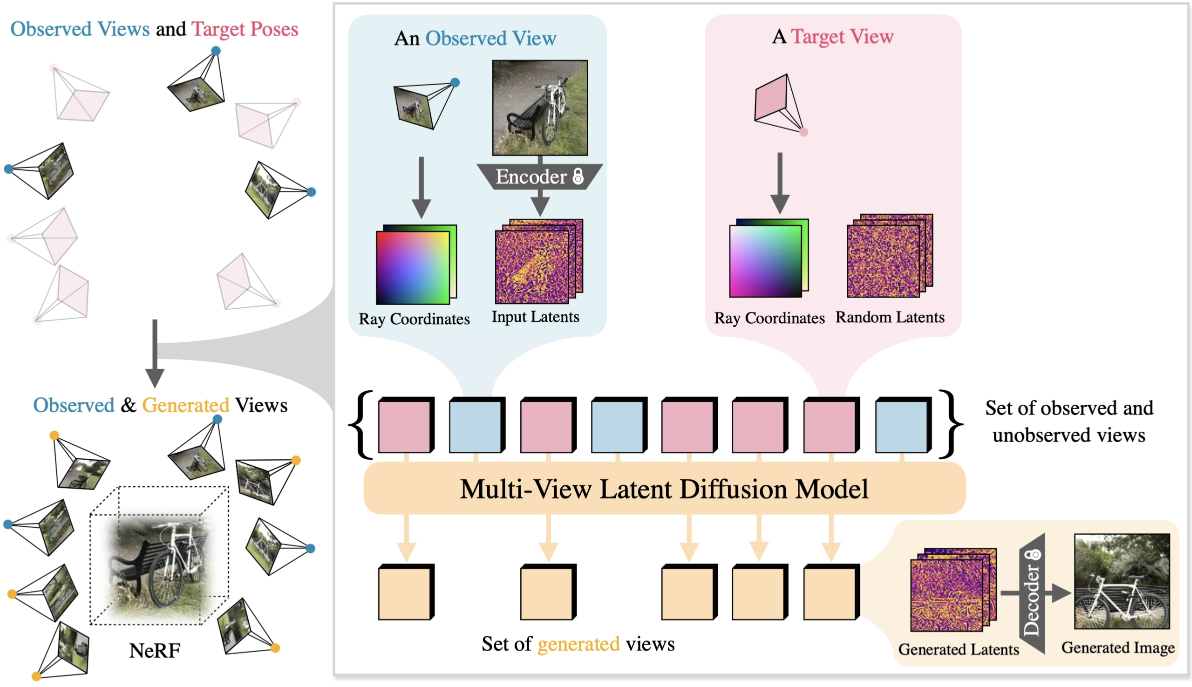

Figure 2: The pipeline of CAT3D

generation (∼5s for 80 views) through efficient parallel sampling. By generating multiple consistent views, these methods provide robust input for traditional 3D reconstruction pipelines, allowing them to better handle challenging sparse-view scenarios. Compared with transformer-based feed-forward methods, these approaches can effectively handle scene-level reconstruction from sparse views, dealing with complex geometry and varying appearance.

Limitations These models struggle with varying camera intrinsics across different views, limiting their applicability to multi-camera setups. While the consistency has been greatly improved, the generated views may still contain minor inconsistencies that can affect final 3D reconstruction quality. Moreover, the effectiveness depends heavily on carefully designed camera trajectories, which can be challenging to determine automatically for complex environments.

3.2.3 Neural Matching Models

Neural matching-based methods, represented by DUSt3R (Wang et al., 2024c), introduce a paradigm shift in 3D vision by predicting pointmaps - dense 2D fields of 3D points that form a one-to-one mapping between image pixels and scene points ($I_{i,j} \leftrightarrow X_{i,j}$), without requiring any prior camera calibration or pose information. The key innovation lies in expressing multiple pointmaps in a common canonical coordinate frame, enabling direct dense reconstruction from unconstrained image collections. Unlike traditional pipelines that rely on sequential steps (feature matching, pose estimation, then dense reconstruction), DUSt3R unifies these tasks through end-to-end learning using a transformer-based architecture. This COLMAP-free property makes 3D reconstruction more accessible and has motivated several subsequent works. For example, Splatt3R (Ye et al., 2024), NoPoSplat (Ye et al., 2024) and PF3plat (Hong et al., 2024a) propose to train a Gaussian head to directly predict pixel-aligned 3D Gaussians. In this way, they method achieve two-view reconstruction with high-quality novel view synthesis.

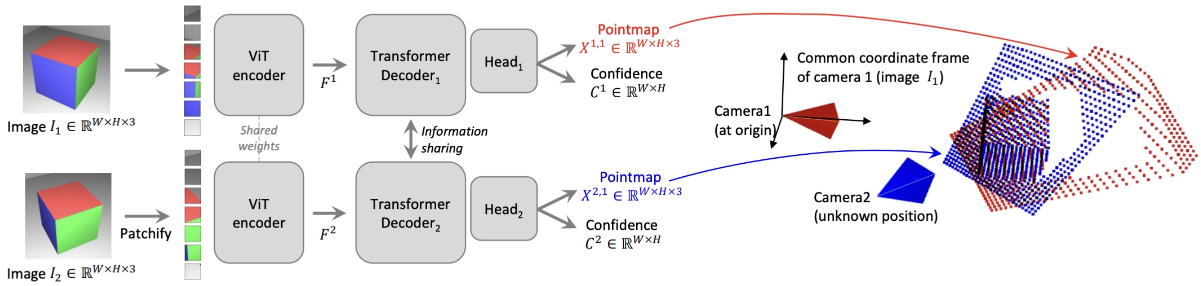
Image $I_1 \in \mathbb{R}^{W\times H\times 3}$
ViT encoder
F^1
Transformer Decoder$_1$
Head$_1$
Pointmap
$X^{1,1} \in \mathbb{R}^{W\times H\times 3}$
Confidence
$C^1 \in \mathbb{R}^{W\times H}$
Shared weights
Information sharing
Patchify
Image $I_2 \in \mathbb{R}^{W\times H\times 3}$
ViT encoder
F^2
Transformer Decoder$_2$
Head$_2$
Pointmap
$X^{2,1} \in \mathbb{R}^{W\times H\times 3}$
Confidence
$C^2 \in \mathbb{R}^{W\times H}$
Common coordinate frame of camera 1 (image I_1)
Camera1 (at origin)
Camera2 (unknown position)

Figure 3: The pipeline of DUSt3R

Architecture Design The core architecture maps an uncalibrated image pair to pointmaps in a canonical coordinate system:

$$f_\theta : \{I_1, I_2\} \mapsto \{X_{1,1}, X_{2,1}, C_1, C_2\} \tag{7}$$

where $X_{i,j} \in \mathbb{R}^{W\times H\times 3}$ represents the pointmap from view i expressed in view j's coordinate frame (aligned with first view), and $C_i \in \mathbb{R}^{W\times H}$ is the associated confidence map. For camera intrinsics estimation, since $X_{1,1}$ is expressed in I_1's coordinate frame, the focal length f_1^* can be recovered through:

$$f_1^* = \arg\min_{f_1} \sum_{i=0}^{W} \sum_{j=0}^{H} C_{1,1}^{i,j} \left\| (i', j') - f_1 \frac{(X_{1,1}^{i,j,0}, X_{1,1}^{i,j,1})}{X_{1,1}^{i,j,2}} \right\| \tag{8}$$

where (i', j') are centered pixel coordinates. For relative pose estimation between views, given pointmaps $X_{1,1}$ and $X_{1,2}$, the relative pose $P^* = [R^*|t^*]$ can be obtained through:

$$R^*, t^* = \arg\min_{\alpha,R,t} \sum_i C_{1,1}^i C_{1,2}^i \|\alpha(RX_{1,1}^i + t) - X_{1,2}^i\|^2 \tag{9}$$

Multi-view Processing For N-view reconstruction, DUSt3R first constructs a connectivity graph $G(V, E)$ where images form vertices V and edges E indicate shared visual content. The globally aligned pointmaps $\{\chi_n\}_{n=1}^N$ are then recovered by optimizing:

$$\chi^* = \arg\min_{\chi,P,\alpha} \sum_{e\in E} \sum_{v\in e} \sum_{i=1}^{HW} C_i^{v,e} \|\chi_i^v - \alpha^e P^e X_i^{v,e}\| \tag{10}$$

where P^e and α^e are the pose and scale for each pair e, enforcing $\prod_e \alpha^e = 1$ to avoid trivial solutions.

Key Advantages This point-based approach offers unique benefits: (1) Direct prediction in canonical coordinates eliminates the need for explicit camera poses and intrinsics during reconstruction; (2) The dense point representation enables reliable geometry and camera parameter estimation across wide baselines; (3) The global optimization allows consistent multi-view fusion without traditional bundle adjustment; (4) Equipped with Gaussian head, these methods can directly reconstruct scenes from sparsely collected images, and perform novel view synthesis.

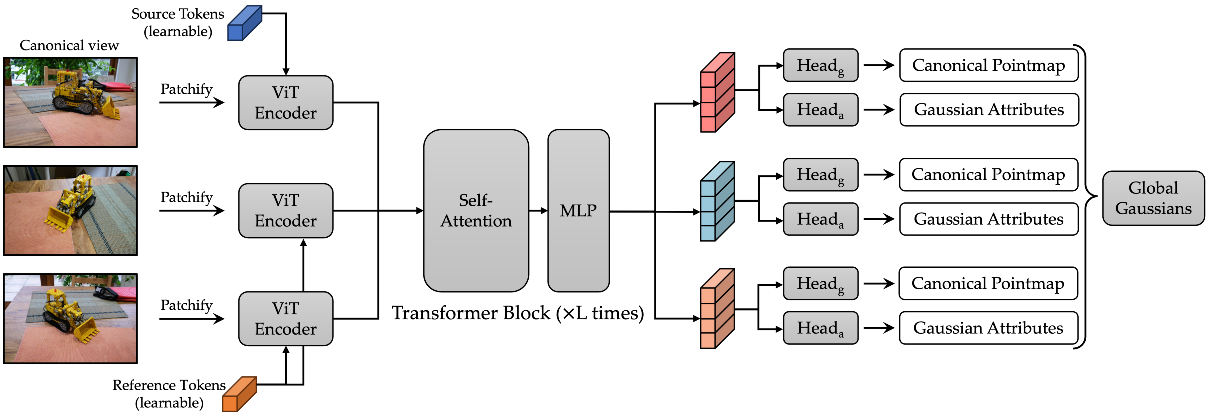

Figure 4: The whole pipeline of my proposal.

Limitations Like all uncalibrated reconstruction methods, these methods suffer from scale ambiguity in the reconstructed geometry. More critically, the pairwise processing design introduces significant scalability issues - for N input images, the method needs to perform $\binom{N}{2}$ pair matches, making it computationally expensive for large image sets. While the subsequent global matching can align these pairwise reconstructions, the lack of bundle adjustment means reconstruction quality degrades noticeably as the number of input views increases, particularly in terms of geometric consistency across distant views. This degradation becomes more pronounced when processing long sequences or large-scale scenes with many input images.

3.3 My Proposal

Based on the analysis of current approaches, achieving high-quality sparse scene reconstruction inevitably involves trade-offs between different desirable properties. Through careful examination of the limitations and advantages of existing methods, I identify three crucial requirements for an effective solution: ***First, the approach should be pose-free***, eliminating the dependency on accurate camera poses that often creates a bottleneck in real-world applications. ***Second, the method should be feed-forward in nature***, directly outputting 3D scene representations without requiring expensive per-scene optimization or iterative refinement. ***Third, the approach must be flexible enough to handle variable numbers of input views*** while maintaining reconstruction quality.

To satisfy these requirements, I propose a novel hybrid approach that combines the strengths of transformer-based feed-forward models (like GS-LRM) and neural matching methods (like DUSt3R). This unified framework aims to leverage both the efficient scene-level understanding capabilities of transformers and the robust pose-free matching mechanism of pointmap-based methods.

3.3.1 Pipeline

The overall pipeline of my proposed approach is illustrated in Fig. 4. My Scene Reconstruction Model (SRM) consists of three main components: Image Tokenization and View Encoding; ViT Decoder; and Gaussian Parameter Prediction Heads. *This proposal presents my preliminary design adapted from prior researches. The network architecture requires further optimization and validation to meet task requirements.*

Image Tokenization and View Encoding In the first stage of my pipeline, I focus on transforming input images into latent representations optimized for multi-view reasoning. I utilize a ViT encoder pre-trained on CroCo (Weinzaepfel et al., 2022) to convert each input image into token sequences through shared-weight encoding. To establish a consistent reconstruction reference frame, I introduce learnable source and reference tokens that define a canonical space for scene representation, which is inspired by Wang et al. (2024a); Peebles and Xie (2023). These view encoding vectors enable reconstruction relative to a fixed source view while supporting pose estimation across arbitrary input views.

ViT Decoder The tokenized representations from multiple views are concatenated into a unified sequence and processed by a ViT decoder comprising L transformer blocks. Each block consists of a self-attention layer that models relationships between all tokens, followed by a MLP. To facilitate effective multi-view reasoning and geometric consistency, SRM incorporates cross-attention mechanisms between different views. Specifically, these cross-attention layers enable each view's tokens to attend to and aggregate information from other views, capturing both local correspondences and global scene structure. At the decoder's output, the processed features are partitioned into view-specific representations, with each partition corresponding to one input frame while maintaining the learned geometric relationships established through the attention mechanisms.

Gaussian Parameter Prediction Heads To generate 3D representations for high-quality novel view synthesis, SRM employs two DPT-based (Ranftl et al., 2021) prediction heads. The first head predicts Gaussian center positions using transformer decoder features exclusively, following DUSt3R and MASt3R. The second head predicts remaining Gaussian parameters using both ViT decoder features and direct RGB input. This RGB shortcut enables efficient texture information flow, which is critical for preserving fine details in the 3D reconstruction.

4 Current Research Matching

I have worked in 3D reconstruction for several years, with 6 first authored papers published in top venues. My track record in developing practical 3D vision solutions (*e.g.*, **GaussianObject with 900+ GitHub stars**) demonstrates my ability to deliver impactful research outcomes in this direction. I show some related experience as follows:

Sparse Object Reconstruction As the lead author of GaussianObject (Accepted by SIGGRAPH Asia 2024), I developed techniques for high-quality 3D reconstruction from as few as 4 images. By introducing visual hull constraints and floater elimination, the method builds reliable multi-view consistency even with extremely sparse inputs. To address the information compression, I propose to finetune a Control-Net with self-feeding manner and then distill the rich object prior to the reconstruction of target object.

Single Image to 3D Scene In LiftImage3D, I explored leveraging video diffusion models' generative prior for single-image 3D reconstruction. The work proposes a distortion-aware 3D Gaussian representation using hexplanes to jointly model 3D geometry and generation-induced distortions.

Pose-Free Reconstruction I have developed a COLMAP-free variant in GaussianObject that achieves competitive quality without requiring pre-given camera poses. This experience with pose-free reconstruction through neural matching directly supports our proposed neural matching approach.

The proposed research addresses critical challenges in 3D vision that impact applications from autonomous driving to augmented reality. As these technologies become increasingly vital, developing robust and efficient 3D understanding systems is essential. My expertise in neural matching, multi-view geometry, and generative models, demonstrated through high-impact publications and widely-adopted open-source work, positions me ideally to advance this important field through novel, practical solutions that bridge current theoretical and applied gaps.

References

Burgess, J., Wang, K.-C., and Yeung, S.: Viewpoint Textual Inversion: Unleashing Novel View Synthesis with Pretrained 2D Diffusion Models, ECCV, 2024.

Chan, E. R., Nagano, K., Chan, M. A., Bergman, A. W., Park, J. J., Levy, A., Aittala, M., De Mello, S., Karras, T., and Wetzstein, G.: GeNVS: Generative novel view synthesis with 3D-aware diffusion models, 2023.

Chen, R., Chen, Y., Jiao, N., and Jia, K.: Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22 246–22 256, 2023.

Deng, K., Liu, A., Zhu, J.-Y., and Ramanan, D.: Depth-supervised NeRF: Fewer Views and Faster Training for Free, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 12872–12881, https://doi.org/10.1109/CVPR52688.2022.01254, https://ieeexplore.ieee.org/document/9880067, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations, https://openreview.net/forum?id=YicbFdNTTy, 2021.

Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J. T., and Poole, B.: CAT3D: Create Anything in 3D with Multi-View Diffusion Models, arXiv preprint arXiv:2405.10314, 2024.

Guangcong, Chen, Z., Loy, C. C., and Liu, Z.: SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis, IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

Haque, A., Tancik, M., Efros, A., Holynski, A., and Kanazawa, A.: Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

Hong, S., Jung, J., Shin, H., Han, J., Yang, J., Luo, C., and Kim, S.: PF3plat: Pose-Free Feed-Forward 3D Gaussian Splatting, arXiv preprint arXiv:2410.22128, 2024a.

Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., and Tan, H.: Lrm: Large reconstruction model for single image to 3d, ICLR, 2024b.

Jain, A., Tancik, M., and Abbeel, P.: Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), p. 5865–5874, https://doi.org/10.1109/ICCV48922.2021.00583, https://ieeexplore.ieee.org/document/9710250, 2021.

Jang, W. and Agapito, L.: NViST: In the Wild New View Synthesis from a Single Image with Transformers, CVPR, 2024.

Jiang, H., Jiang, Z., Zhao, Y., and Huang, Q.: LEAP: Liberate Sparse-view 3D Modeling from Camera Poses, ICLR, 2024.

Kim, M., Seo, S., and Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering, in: CVPR, pp. 12 912–12 921, 2022.

Leroy, V., Cabon, Y., and Revaud, J.: Grounding Image Matching in 3D with MASt3R, arXiv preprint arXiv:2406.09756, 2024.

Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., and Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model, ICLR, 2024.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y.: Magic3D: High-Resolution Text-to-3D Content Creation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C.: Zero-1-to-3: Zero-shot One Image to 3D Object, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9298–9309, 2023a.

Liu, X., Kao, S.-h., Chen, J., Tai, Y.-W., and Tang, C.-K.: Deceptive-NeRF: Enhancing NeRF Reconstruction using Pseudo-Observations from Diffusion Models, arXiv preprint arXiv:2305.15171, 2023b.

Metzer, G., Richardson, E., Patashnik, O., Giryes, R., and Cohen-Or, D.: Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12 663–12 673, https://doi.org/10.1109/CVPR52729.2023.01218, 2023.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, in: ECCV, vol. 12346 of *Lecture Notes in Computer Science*, pp. 405–421, Springer, https://doi.org/10.1007/978-3-030-58452-8_24, https://doi.org/10.1007/978-3-030-58452-8_24, 2020.

Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulò, S. R., Nießner, M., and Kontschieder, P.: MultiDiff: Consistent Novel View Synthesis from a Single Image, in: CVPR, pp. 10 258–10 268, 2024.

Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S. M., Geiger, A., and Radwan, N.: RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 5470–5480, https://doi.org/10.1109/CVPR52688.2022.00540, https://ieeexplore.ieee.org/document/9879664, 2022.

Pan, Z., Yang, Z., Zhu, X., and Zhang, L.: Fast Dynamic 3D Object Generation from a Single-view Video, arXiv preprint arXiv 2401.08742, 2024.

Peebles, W. and Xie, S.: Scalable diffusion models with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion, ICLR, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th International Conference on Machine Learning, p. 8748–8763, PMLR, https://proceedings.mlr.press/v139/radford21a.html, 2021.

Ranftl, R., Bochkovskiy, A., and Koltun, V.: Vision Transformers for Dense Prediction, ICCV, 2021.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V.: Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44, 2022.

Roessle, B., Barron, J. T., Mildenhall, B., Srinivasan, P. P., and Nießner, M.: Dense depth priors for neural radiance fields from sparse input views, in: CVPR, pp. 12 892–12 901, 2022.

Schönberger, J. L. and Frahm, J.-M.: Structure-from-Motion Revisited, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Seo, S., Chang, Y., and Kwak, N.: Flipnerf: Flipped reflection rays for few-shot novel view synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22 883–22 893, 2023.

Shao, R., Sun, J., Peng, C., Zheng, Z., Zhou, B., Zhang, H., and Liu, Y.: Control4D: Efficient 4D Portrait Editing with Text, CVPR, 2024.

Shi, R., Wei, X., Wang, C., and Su, H.: ZeroRF: Fast Sparse View 360° Reconstruction with Zero Pretraining, CVPR, 2024a.

Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., and Yang, X.: MVDream: Multi-view Diffusion for 3D Generation, ICLR, 2024b.

Smart, B., Zheng, C., Laina, I., and Prisacariu, V. A.: Splatt3r: Zero-shot gaussian splatting from uncalibarated image pairs, arXiv preprint arXiv:2408.13912, 2024.

Somraj, N. and Soundararajan, R.: ViP-NeRF: Visibility Prior for Sparse Input Neural Radiance Fields, in: ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), https://doi.org/10.1145/3588432.3591539, 2023.

Somraj, N., Karanayil, A., and Soundararajan, R.: SimpleNeRF: Regularizing Sparse Input Neural Radiance Fields with Simpler Solutions, in: SIGGRAPH Asia 2023 Conference Papers, SA '23, p. 1–11, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3610548.3618188, https://dl.acm.org/doi/10.1145/3610548.3618188, 2023.

Somraj, N., Karanayil, A., Mupparaju, S. H., and Soundararajan, R.: Simple-RF: Regularizing Sparse Input Radiance Fields with Simpler Solutions, arXiv preprint arXiv:2404.19015, 2024.

Song, J., Park, S., An, H., Cho, S., Kwak, M.-S., Cho, S., and Kim, S.: DaRF: Boosting Radiance Fields from Sparse Inputs with Monocular Depth Adaptation, 2023 NIPS, 2023a.

Song, L., Li, Z., Gong, X., Chen, L., Chen, Z., Xu, Y., and Yuan, J.: Harnessing Low-Frequency Neural Fields for Few-Shot View Synthesis, arXiv preprint arXiv:2303.08370, 2023b.

Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., and Liu, Z.: LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation, ECCV, 2024a.

Tang, J., Ren, J., Zhou, H., Liu, Z., and Zeng, G.: DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation, ICLR, 2024b.

Wang, H., Du, X., Li, J., Yeh, R. A., and Shakhnarovich, G.: Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12 619–12 629, https://doi.org/10.1109/CVPR52729.2023.01214, 2023a.

Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., and Zhang, K.: PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction, ICLR, 2024a.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., and Revaud, J.: DUSt3R: Geometric 3D Vision Made Easy, CVPR, 2024b.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., and Revaud, J.: Dust3r: Geometric 3d vision made easy, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20 697–20 709, 2024c.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J.: ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation, in: Advances in Neural Information Processing Systems (NeurIPS), 2023b.

Wei, X., Zhang, K., Bi, S., Tan, H., Luan, F., Deschaintre, V., Sunkavalli, K., Su, H., and Xu, Z.: MeshLRM: Large Reconstruction Model for High-Quality Mesh, arXiv preprint arXiv:2404.12385, 2024.

Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., and Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion, Advances in Neural Information Processing Systems, 35, 3502–3516, 2022.

Weng, Z., Liu, J., Tan, H., Xu, Z., Zhou, Y., Yeung-Levy, S., and Yang, J.: Template-Free Single-View 3D Human Digitalization with Diffusion-Guided LRM, Preprint, 2023.

Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P. P., Verbin, D., Barron, J. T., Poole, B., et al.: ReconFusion: 3D Reconstruction with Diffusion Priors, CVPR, 2024.

Wynn, J. and Turmukhambetov, D.: DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 4180–4189, https://doi.org/10.1109/CVPR52729.2023.00407, https://ieeexplore.ieee.org/document/10203756, 2023.

Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H., and Wang, Z.: SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image, in: Computer Vision – ECCV 2022, edited by Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., Lecture Notes in Computer Science, p. 736–753, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-20047-2_42, 2022.

Xu, D., Yuan, Y., Mardani, M., Liu, S., Song, J., Wang, Z., and Vahdat, A.: AGG: Amortized Generative 3D Gaussians for Single Image to 3D, arXiv preprint 2401.04099, 2024a.

Xu, Y., Shi, Z., Yifan, W., Chen, H., Yang, C., Peng, S., Shen, Y., and Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation, ECCV, 2024b.

Xu, Y., Tan, H., Luan, F., Bi, S., Wang, P., Li, J., Shi, Z., Sunkavalli, K., Wetzstein, G., Xu, Z., and Zhang, K.: DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model, ICLR, 2024c.

Yang, C., Li, S., Fang, J., Liang, R., Xie, L., Zhang, X., Shen, W., and Tian, Q.: GaussianObject: High-Quality 3D Object Reconstruction from Four Views with Gaussian Splatting, ACM Transactions on Graphics, 2024.

Yang, J., Pavone, M., and Wang, Y.: FreeNeRF: Improving Few-Shot Neural Rendering with Free Frequency Regularization, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 8254–8263, https://doi.org/10.1109/CVPR52729.2023.00798, https://ieeexplore.ieee.org/document/10205351, 2023.

Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.-H., and Peng, S.: No Pose, No Problem: Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images, arXiv preprint arXiv:2410.24207, 2024.

Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., and Wang, X.: GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models, CVPR, 2024.

Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.-T., Shan, Y., and Tian, Y.: ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis, arXiv preprint arXiv:2409.02048, 2024.

Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., and Xu, Z.: GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting, arXiv, 2024.

Zhang, L., Rao, A., and Agrawala, M.: Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847, 2023.

Zhou, Z. and Tulsiani, S.: SparseFusion: Distilling View-Conditioned Diffusion for 3D Reconstruction, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 12588–12597, https://doi.org/10.1109/CVPR52729.2023.01211, https://ieeexplore.ieee.org/document/10204403, 2023.

Zhu, J. and Zhuang, P.: HiFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance, ICLR, 2024.

Ziwen, C., Tan, H., Zhang, K., Bi, S., Luan, F., Hong, Y., Fuxin, L., and Xu, Z.: Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats, arXiv preprint arXiv:2410.12781, 2024.

Zou, Z.-X., Yu, Z., Guo, Y.-C., Li, Y., Liang, D., Cao, Y.-P., and Zhang, S.-H.: Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers, CVPR, 2024.