Advances in Feed-Forward 3D Reconstruction and View Synthesis

Jiahui Zhang, Yuelei Li, Anpei Chen, Muyu Xu, Kunhao Liu, Jianyuan Wang, Xiaoxiao Long, Hanxue Liang, Zexiang Xu, Hao Su, Christian Theobalt, Christian Rupprecht, Andrea Vedaldi, Hanspeter Pfister, Shijian Lu§, Fangneng Zhan

Abstract—3D reconstruction and view synthesis are foundational problems in computer vision, graphics, and immersive technologies such as augmented reality (AR), virtual reality (VR), and digital twins. Traditional methods rely on computationally intensive iterative optimization in a complex chain, limiting their applicability in real-world scenarios. Recent advances in feed-forward approaches, driven by deep learning, have revolutionized this field by enabling fast and generalizable 3D reconstruction and view synthesis. This survey offers a comprehensive review of feed-forward techniques for 3D reconstruction and view synthesis, with a taxonomy according to the underlying representation architectures including point cloud, 3D Gaussian Splatting (3DGS), Neural Radiance Fields (NeRF), etc. We examine key tasks such as pose-free reconstruction, dynamic 3D reconstruction, and 3D-aware image and video synthesis, highlighting their applications in digital humans, SLAM, robotics, and beyond. In addition, we review commonly used datasets with detailed statistics, along with evaluation protocols for various downstream tasks. We conclude by discussing open research challenges and promising directions for future work, emphasizing the potential of feed-forward approaches to advance the state of the art in 3D vision. A project page associated with this survey is available at Feed-Forward-3D.

Index Terms—Feed-forward Model, 3D Reconstruction, Neural Rendering, Radiance Fields, NeRF, 3DGS.

1 Introduction

3D reconstruction and rendering are long-standing and central challenges in computer vision and computer graphics. They enable a wide range of applications, from digital content creation, augmented reality, and virtual reality to robotics, autonomous systems, and digital twins. Traditionally, high-quality 3D reconstruction has relied on optimization-based pipelines such as Structure-from-Motion (SfM), Multi-View Stereo (MVS), and differentiable rendering. While effective in controlled scenarios, these methods are often computationally expensive, slow to converge, and dependent on large-scale or precisely calibrated datasets—limiting their practicality in real-time or open-world environments.

In recent years, fueled by breakthroughs in deep learning and neural representations, *feed-forward 3D reconstruction and view synthesis* have emerged as a transformative alternative as shown in Fig. 1. Unlike classical methods that require iterative optimization per scene, feed-forward models infer 3D geometry or novel views in a single forward pass, enabling orders-of-magnitude faster inference with improved generalization. These models exploit learned priors and large-scale training to predict 3D structure, rendering outputs, or both—directly from unposed, sparse, or monocular inputs—making them especially appealing for time-sensitive and scalable applications such as interactive graphics, robotic perception, and generative media.

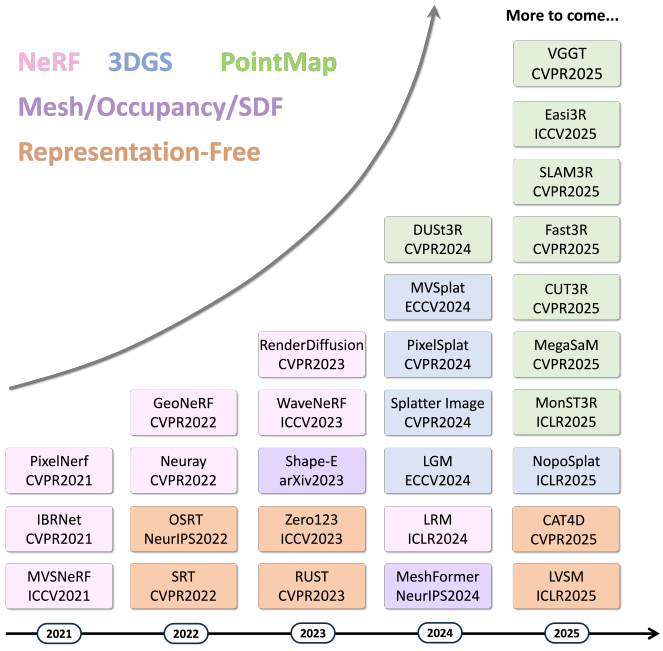

Fig. 1: A summary of the representative works based on their category and timeline.

- *Jiahui Zhang, Muyu Xu, Kunhao Liu, and Shijian Lu are with the Nanyang Technological University, Singapore.*
- *Yuelei Li is with the California Institute of Technology, USA.*
- *Anpei Chen is with the Westlake University, China.*
- *Jianyuan Wang, Christian Rupprecht, and Andrea Vedaldi are with the University of Oxford, UK.*
- *Xiaoxiao Long is with the Nanjing University, China.*
- *Hanxue Liang is with the University of Cambridge, UK.*
- *Zexiang Xu is with the Hillbot, USA.*
- *Hao Su is with the University of California, San Diego, USA.*
- *Christian Theobalt is with the Max Planck Institute for Informatics, Germany.*
- *Hanspeter Pfister is with the Harvard University, USA.*
- *Fangneng Zhan is with the Harvard University, USA, and the Massachusetts Institute of Technology, USA.*
- *§ denotes corresponding author, E-mail: shijian.lu@ntu.edu.sg.*

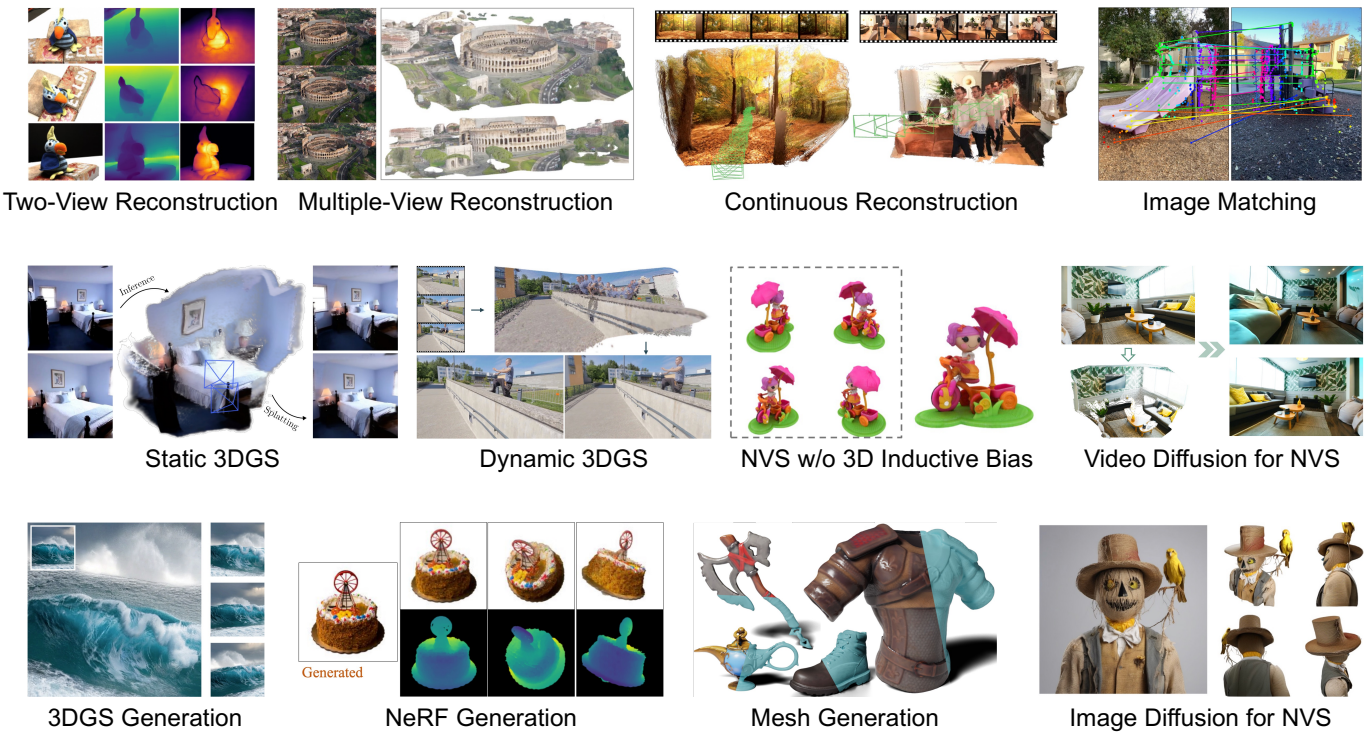

Fig. 2: This state-of-the-art survey discusses the methodology and application of feed-forward models for various 3D reconstruction and Novel View Synthesis (NVS) tasks. The samples are adapted from [1]–[12].

This survey presents a comprehensive review of feed-forward methods for 3D reconstruction and view synthesis, with an emphasis on the **core architectures, scene representations, and downstream applications** that define this fast-evolving area. We systematically categorize existing approaches based on their underlying scene representations, which fundamentally determine how 3D structure and appearance are modeled and rendered. Specifically, we identify five major categories: models built on **Neural Radiance Fields (NeRF)** [13], which leverage volumetric rendering through learned radiance fields; **pointmap-based** approaches [1], which operate on pixel-aligned 3D pointmaps; **3D Gaussian Splatting (3DGS)**-based models [14], which use rasterizable Gaussian primitives for fast, high-quality rendering; **mesh-, occupancy-, and signed distance function (SDF)**-based methods, which rely on explicit or implicit scene representations; and **representation-free** models, which leverage deep neural networks to synthesize views directly without a explicit 3D representation. For each category, we provide an in-depth analysis of representative and state-of-the-art methods, examining their core architectural designs, feature representations, and the inductive biases embedded in their formulations. We emphasize the key design principles that govern their effectiveness—such as spatial reasoning, multiview consistency, and geometric priors—and elaborate how subsequent works extend, refine, or generalize these ideas within the research lines.

We also discuss high-impact 3D vision applications enabled by feed-forward methods as shown in Fig. 2, which provide scalable, fast, and generalizable solutions across domains. These include pose-free and dynamic 3D reconstruction, 3D-aware image and video synthesis, and camera-controllable video generation. Additionally, these models facilitate semantic reasoning and dense matching, advancing tasks such as 3D-aware segmentation, optical flow estimation. In robotics and SLAM, feed-forward models enable real-time scene understanding and tracking, while in digital humans, they support efficient yet generalizable avatar reconstruction from sparse inputs.

To facilitate future progress, we review widely used benchmark datasets and evaluation protocols for feed-forward 3D reconstruction and view synthesis. These datasets cover synthetic and real-world scenes across objects, indoor and outdoor environments, and static or dynamic settings, with varying levels of annotation such as RGB, depth, LiDAR, and optical flow. We also summarize standard evaluation metrics for assessing image quality, geometry accuracy, camera pose estimation, etc. Together, these benchmarks and metrics provide essential foundations for comparing methods, diagnosing failure cases, and driving progress toward more generalizable, accurate, and robust 3D models.

Despite impressive progress, feed-forward models still face major challenges, including limited modality diversity in datasets, poor generalization in free-viewpoint synthesis, and the high computational cost of long-context processing. Addressing these issues will require advances in representation design, efficient architectures, and scalable supervision. Finally, we conclude with the societal impact of this technology, highlighting the importance of responsible deployment and transparent modeling practices.

2 METHODS

In the following, we broadly categorize the feed-forward 3D reconstruction and view synthesis methods into five categories based on their underlying representation: NeRF models (Sec. 2.1), Pointmap models (Sec. 2.2), 3DGS models (Sec. 2.3), models employing other common representations (*e.g.*, mesh, occupancy, SDFs in Sec. 2.4), and 3D representation-free models (Sec. 2.5).

2.1 NeRF

Neural radiance fields (NeRF) [13] have recently gained significant attention for high-quality novel view synthesis using implicit scene representations and differentiable volume rendering. By leveraging MLPs, NeRF reconstructs 3D scenes from multi-view 2D images, enabling the generation of novel views with excellent multi-view consistency. However, a major limitation of NeRF is its requirement for per-scene optimization, which restricts its generalization to unseen scenes. To address this, feed-forward approaches have been proposed, where neural networks learn to infer NeRF representations directly from sparse input views, thereby eliminating the need for scene-specific optimization. As a pioneering feed-forward NeRF work, PixelNeRF [18] introduces a conditional NeRF framework that leverages pixel-aligned image features extracted from input images, allowing the model to generalize across diverse scenes and perform novel view synthesis from sparse observations. A large number of follow-ups adopt various techniques for feed-forward NeRF and we broadly categorize these methods into the following categories based on feature representations.

2.1.1 1D Feature-based Methods

Several methods have been proposed to encode a global 1D latent code for NeRF prediction, where the same latent code is shared across all 3D points in a scene. For example, CodeNeRF [15], as illustrated in Fig. 3 (a), introduces a disentanglement strategy that jointly learns separate embeddings for texture and shape, along with an MLP conditioned on these embeddings to predict the color and volumetric density of each 3D point. ShaRF [19] introduces latent codes for shape and appearance, which serve as conditioning inputs for NeRF reconstruction. These codes enable control over the geometry and visual appearance of the reconstructed 3D objects. Besides, Shap-E [20], following Point-E [21], encodes point clouds and RGBA input images into a series of latent vectors which are subsequently utilized for NeRF prediction.

2.1.2 2D Feature-based Methods

2D feature-based methods typically leverage an image encoder to extract image features of source views and obtain features of arbitrary 3D points by ray projection. For example, GRF [16], as illustrated in Fig. 3(b) projects each 3D point along a camera ray onto source views to extract corresponding multi-view features. These features are then aggregated and passed through an MLP to predict RGB color and volumetric density. IBRNet [22] follows a comparable approach, projecting 3D points onto nearby source views to extract image features that are aggregated across views for radiance field inference. NeRFormer [23] also employs ray-projected features and performs multi-view feature aggregation to guide NeRF prediction. Besides, SRF [24] projects 3D points onto multi-view input images to construct a stereo feature matrix, which is processed by a 2D CNN to produce view-aligned features for color and density prediction. To provide additional geometric cues for color and density prediction, GNT [25] introduces a view transformer that leverages epipolar constraint to aggregate projected features from multiple views in a geometrically consistent manner. MatchNeRF [26] explicitly models correspondence information by computing the similarity between ray-projected features from pairs of nearby source views, using this information as a conditioning input for the prediction. ContraNeRF [27] introduces geometry-aware feature extraction and contrastive learning [28] to query features from multiple source views and aggregate them to obtain geometrically enhanced feature maps.

2.1.3 3D Feature-based Methods

3D Volume Features. MVSNeRF [17] is inspired by multi-view stereo (MVS) [29]–[31] and constructs cost volumes from input images as shown in Fig. 3(c). These cost volumes are used to generate a neural scene encoding volume that stores per-voxel features capturing both local geometry and appearance. For any 3D point, its features are obtained via trilinear interpolation from the encoding volume and then decoded by an MLP to predict the corresponding density and color. To enhance rendering quality in both fine-detail areas and occluded regions, GeoNeRF [32] extends MVSNeRF by first constructing cascaded cost volumes for each source view, followed by an attention-based volume aggregation across views. This design enables high-resolution detail reconstruction while effectively handling occlusions. NeuRay [33] is also proposed to address the issue of occlusion. Specifically, it leverages constructed cost volumes to predict the visibility of 3D points, allowing the model to identify feature inconsistencies caused by occlusion and enhance rendering quality in complex scenes with severe self-occlusions. WaveNeRF [34] designs a wavelet multiview stereo that incorporates wavelet frequency volumes into the MVS to preserve high-frequency information and achieve desirable scene geometry reconstruction. Besides, for efficient rendering, ENeRF [35] proposes sampling a limited number of points near the scene surface by predicting the coarse scene geometry from constructed cascade cost volume, enabling improved rendering speed. MuRF [36] eliminates the use of cost volumes for pre-defined reference input views, instead constructing a target view frustum volume to effectively aggregate information from the input images, particularly in scenes with limited overlap between the reference and target views. To improve the quality of geometry estimation, GeFu [37] introduces an adaptive cost aggregation module that reweights the contributions of different source views, allowing the model to learn adaptive weights for constructing cost volumes.

3D Triplane Features. As a pioneering triplane-based method, Large Reconstruction Model (LRM) [10] employs a large transformer-based encoder-decoder architecture and directly regresses a feature triplane representation as shown in Fig. 3(d), enabling NeRF prediction from triplane features.

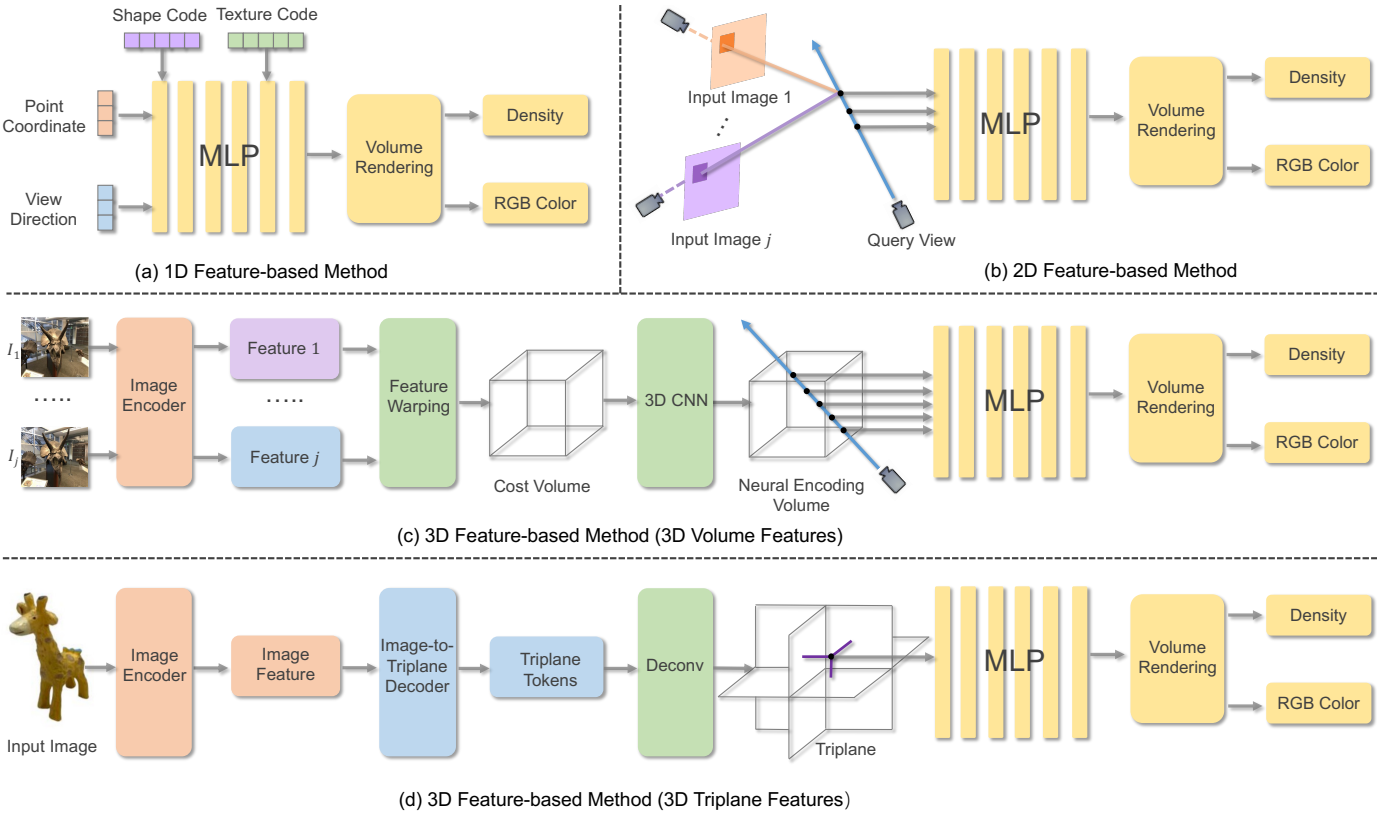

Fig. 3: Typical frameworks of feed-forward NeRF. The samples are adapted from [15] [16] [17] [10].

Pf-LRM [38] extends the LRM into the pose-free setting, which jointly reconstructs the triplane NeRF representations and predicts relative camera poses in a feed-forward manner. TripoSR [39] further enhances LRM through improvements in data curation and rendering, model architecture, and training strategies. Considering the scarcity, licensing constraints, and inherent biases of 3D data, LRM-Zero [40] is proposed to enable training solely on synthesized data from Zeroverse [40]. Besides, several methods combines the large reconstruction model with diffusion model. For example, Instant3D [41] first leverages fine-tuned 2D diffusion model [42] to generate 4-view images from a text prompt and then utilizes a transformer-based large reconstruction model to predict NeRF model. DMV3D [43], inspired by RenderDiffusion [44], incorporates LRM into multi-view diffusion, which gradually reconstructs clean triplane NeRF representation from noisy multi-view images in the diffusion process.

2.1.4 Other Methods.

In addition to the aforementioned methods, several efforts have also focused on feed-forward NeRF reconstruction with other types of features. For example, VisionNeRF [45] proposes to leverage vision transformer [46] and convolutional networks to extract global 1D features and 2D image features respectively, and constructs a multi-level feature maps that serves as the conditioning inputs of NeRF prediction to enhance rendering quality, particularly in occluded regions. MINE [47] integrates NeRF and multiplane image (MPI) [48] representations to enable generalizable, occlusion-aware 3D reconstruction from a single image.

2.2 Pointmap

Pointmaps [1], [49]–[52], encode scene geometry, pixel-to-scene correspondences, and viewpoint relationships, allowing for camera poses, depths and explicit 3D primitive estimation. The pioneering feed-forward pointmap reconstruction method DUSt3R [1] learns a transformer-based encoder-decoder to directly output two pixel-aligned pointmaps from image pairs without posed cameras, enabling dense unconstrained stereo 3D reconstruction. The follow-up work, MASt3R [4], improves DUSt3R by introducing local feature matching. To handle more views, Fast3R [53] builds upon the DUSt3R framework and designs a global fusion transformer to process multi-view inputs simultaneously, eliminating the sequential reconstruction in Spann3R and significantly enhancing the reconstruction quality. SLAM3R [54] introduces an Image-to-Points module that enables simultaneous processing of multi-view inputs, effectively enhancing reconstruction quality without sequential reconstruction. VGGT [2] presents a large feed-forward transformer-based architecture that directly predicts all essential 3D attributes—such as camera intrinsics and extrinsic, point maps, depth maps, and 3D point tracks—without the need for post-processing, leading to the state of the art 3D point and camera pose reconstruction. To reduce memory usage, a couple of workers incrementally process the input and add points to a canonical 3D space by maintaining and incrementally updating a scene's latent state. Spann3R [55] introduces a spatial memory network, enabling multi-view input and improving efficiency to eliminate the need for global alignment. Closely, CUT3R [3] pro-

poses a Continuous Updating Transformer that simultaneously updates the state with new information and retrieves the information stored in the state. This formulation is general and able to handle both videos and photo collections, and processing both static and dynamic scenes. To facilitate accurate 3D reconstruction, Pow3R [56] flexibly integrates available priors at test time—such as camera intrinsics, sparse or dense depth, or relative poses—as lightweight and diverse conditioning. In contrast, Rig3R [57] exploits rig metadata as conditions to improve both camera pose estimation and 3D reconstruction. Besides, several methods are proposed to develop new SfM pipelines to achieve efficient 3D reconstruction. Light3R-SfM [58] replaces traditional optimization-based global alignment with a learnable latent alignment module, enabling efficient SfM technique and 3D reconstruction. Regist3R [59] introduces a stereo foundation model to build a scalable incremental SfM pipeline for efficient 3D reconstruction.

2.3 3DGS

3D Gaussian Splatting (GS) [14] is a recent advance for efficient 3D reconstruction and rendering built upon rasterization. 3DGS and 2DGS are point-based representations that each point associated with geometry attributes (*i.e.*center position, shape, orientation and opacity α) and Spherical Harmonics (SH) appearance attributes. Despite its high fidelity in reconstruction, 3DGS requires per-scene optimization, which limits its training efficiency and generalization capabilities. Recently, feed-forward 3DGS reconstruction methods have emerged, leveraging neural networks to directly predict Gaussian parameters. These approaches eliminate the need for per-scene optimization and enable generalizable novel view synthesis. We categorize these methods based on the representation of predicted Gaussian outputs: image, volume, triplane and pointmap.

2.3.1 Gaussian Image

Gaussian image refers to single-view or multi-view images formed by predicted 3D Gaussians, where each pixel represents a 3D Gaussians. It is a typical and widely used Gaussian representation, and numerous methods adopting this approach will be introduced below.

As illustrated in Fig. 4(a), Splatter Image [60], a pioneering feed-forward 3DGS method, employs a U-Net encoder-decoder architecture [64] to predict pixel-aligned 3D Gaussians, forming Gaussian image for impressive single-view 3D object reconstruction. Flash3D [65] extends this approach to single-view scene-level reconstruction using a similar U-Net encoder-decoder framework.

To improve the reconstruction quality, several methods are subsequently proposed to leverage large models with strong capacity to learn generic scene priors from large-scale datasets for 3D scene reconstruction. Based on 3D large reconstruction model (LRM) [10] that achieves impressive sparse-view 3D object reconstruction by learning general reconstruction priors from extensive datasets of 3D objects, GRM [66] directly maps input image pixels to a set of pixel-aligned 3D Gaussians for feed-forward 3DGS-based object reconstruction. Concurrently, GS-LRM [67] also incorporates Transformer-based LRM into feed-forward 3D Gaussian primitive prediction, which treats per-pixel Gaussian prediction as a sequence-to-sequence mapping and achieves remarkable performance across both objects and scenes. However, its design is limited to constrained viewing coverage, preventing its application to large-scale real-world reconstructions. To address this limitation, Long-LRM [68] introduces a novel architecture that combines Mamba2 blocks [69] with Transformer layers, establishing the first feed-forward Gaussian solution for wide-coverage scene-level reconstruction. Furthermore, to mitigate the reliance of LRM-based methods on accurate input camera poses, FreeSplatter [70] introduces a transformer-based large reconstruction model that jointly predicts Gaussian images and estimates camera poses directly from multi-view inputs. In addition, several methods have been proposed that do not rely on transformer-based LRMs. For example, to improve the reconstruction in non-overlapping and occluded regions of real-world scenes, eFreeSplat [71] leverages 3D priors learned from a large vision transformer encoder [46] and cross-attention decoder pre-trained on 3D cross-view completion for Gaussian image prediction. LGM [72] first introduces pre-trained diffusion models [12], [73]–[75] to generate multi-view images and then leverages 3D priors in a pretrained asymmetric U-Net based large multi-view Gaussian model to achieve multi-view Gaussian image prediction.

Furthermore, to enhance the geometric quality of 3D scene reconstruction, a large number of methods incorporate geometric designs, such as epipolar and cost volumes.

Epipolar-based Methods. As a representative and pioneering epipolar-based method, PixelSplat [5] leverages an epipolar transformer to resolve the scale ambiguity issue and capture cross-view features. It then estimates a probabilistic depth distribution from the image features and predicts pixel-aligned 3D Gaussians. However, PixelSplat is effective only in regions strongly correlated with the input observations. It struggles in areas of high uncertainty, leading to blurry reconstructions that lack high-frequency details or failed reconstruction in unseen regions. LatentSplat [76] proposes to exploit a generative model to obtain high-quality reconstructions in uncertain areas. It leverages epipolar transformer and a Gaussian sampling head to encode two-view inputs to 3D variational Gaussians and finally uses a lightweight VAE-GAN decoder [77] to generate RGB images of novel views. This approach enables high-quality binocular reconstruction of object-centric scenes with full 360° views.

Besides, several methods are proposed to enable pose-free feed-forward 3DGS. For example, GGRt [78] builds upon PixelSplat and introduces a joint learning framework, which is utilized to jointly optimize the camera poses and the 3D Gaussian prediction, reducing the reliance on ground-truth camera poses.

Cost Volume-based Methods. A key limitation of PixelSplat is the inherent ambiguity and unreliability in mapping image features to depth distributions, resulting in suboptimal geometry reconstruction. For accurate 3D Gaussian reconstruction, MVSplat [79] employs a cost volume representation based on plane sweeping in 3D space for multi-view Gaussian image prediction, where cross-view feature similarities encoded in the cost volume offer valuable ge-

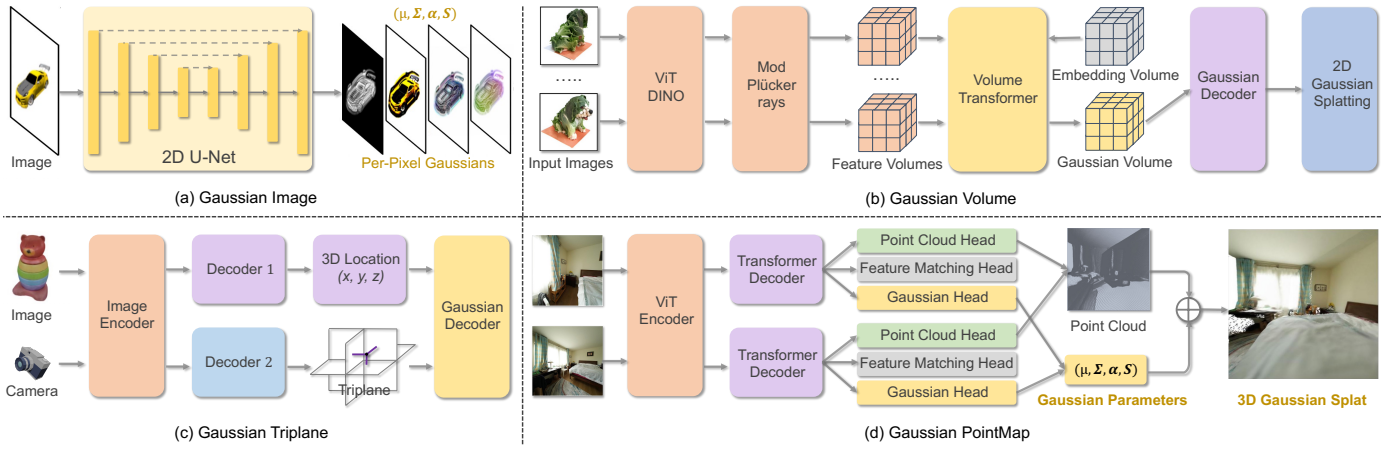

Fig. 4: Typical frameworks for four output 3D Gaussian representations: Gaussian image, Gaussian volume, Gaussian triplane, and Gaussian PointMap. The samples are adapted from [60] [61] [62] [63].

ometric cues for depth estimation. MVSGaussian [80] also employs a cost volume-based pipeline to estimate depth, which is then used for pixel-aligned 3D Gaussian prediction. However, these methods heavily depend on precise multi-view feature matching, which becomes particularly challenging in scenes with occlusions, low texture, or repetitive patterns. To address this issue, TranSplat [81] introduces a depth-aware deformable matching transformer to generate a depth confidence map to enhance the multi-view feature matching, thus improving reconstruction accuracy in areas with low texture or repetitive patterns. Similarly, DepthSplat [82] leverages robust monocular depth estimation to enhance feed-forward 3DGS reconstruction. Specifically, it combines pre-trained monocular depth features with multi-view feature matching, preserving multi-view depth consistency while improving robustness in difficult scenarios such as occlusions, textureless regions, and reflective surfaces. The predicted multi-view depth maps are then utilized to determine the Gaussian centers, while a lightweight network estimates the remaining Gaussian parameters. Another recent work, HiSplat [83], is introduced to address the limitation of feed-forward 3DGS reconstruction in lacking hierarchical representations, which makes it difficult to simultaneously capture large-scale structures and fine texture details. After leveraging cost volume to obtain the depth and Gaussian features, HiSplat creates large, coarse-grained Gaussian image to define the primary structure, then adds finer Gaussians around them to progressively refine and enrich the texture details. PanSplat [84] also investigates hierarchical Gaussian images for 4K panorama view synthesis. It employs a transformer-based network to build a hierarchical spherical cost volume, enabling high-resolution 3D geometry with enhanced efficiency. MVSplat360 [85] extends MVSplat to support 360° novel view synthesis for large-scale real-world scenes. It incorporates MVSplat with latent video diffusion [86], leveraging the 3DGS rendered features as conditioning input for the video diffusion to generate visually compelling novel views with strong multi-view consistency and premise geometry.

Additionally, Pf3plat [87] is proposed to achieve pose-free feed-forward 3DGS reconstruction and rendering. It introduces a coarse-to-fine strategy to estimate the depth, confidence and camera poses and utilize them to perform 3D Gaussian prediction with the constructed multi-stereo cost volume and guidance cost volume.

2.3.2 Gaussian Volume

Gaussian volume [61], [88] represents 3D with Gaussian voxel grids, where each voxel comprises multiple Gaussian primitives. A typical feed-forward 3DGS method using a Gaussian volume representation is LaRa [61], which aims to reduce the heavy training cost associated with 360° bounded radiance field reconstruction. As shown in Fig. 4 (b), it first builds 3D feature and embedding volumes and then leverages a volume transformer to reconstruct a Gaussian volume, enabling progressively and implicitly feature matching and leading to higher quality results and faster convergence. GaussianCube [89] proposes a structured and explicit radiance representation for 3D object generation from a single image.

2.3.3 Gaussian Triplane

Gaussian triplane refers to a hybrid 3D representation that effectively combines the high-quality representation of triplanes with the efficiency of 3D Gaussian splatting. It is typically constructed by triplane-based 3DGS methods, which aim to predict a triplane representation first and then leverage the latent triplane features to decode 3D Gaussians, as illustrated in Fig. 4(c). For example, TriplaneGaussian [62] leverages several transformer-based networks pre-trained in large-scale datasets to build Gaussian triplane, enabling high-quality single-view 3D reconstruction. AGG [90] also mixes triplane and 3D Gaussians, which first represents scene textures as triplane and then decodes 3D Gaussians from triplane-based texture features queried by 3D locations.

2.3.4 Gaussian PointMap

Gaussian pointmap refers to a hybrid 3D representation that combines pointmaps with 3D Gaussians. It is typically constructed by pointmap-based 3DGS methods which aim to generate dense Gaussian pointmaps to enable pose-free sparse-view reconstruction and rendering. Specifically,

these methods often leverage pointmaps as geometric priors, upon which 3D Gaussians are predicted, as illustrated in Fig. 4(d). With the advent of a series of feed-forward pointmap reconstruction methods [1], [4], [55], which regress the dense pointmaps directly from raw unposed images, one research line of pointmap-based methods is that directly leverages the pointmap reconstruction methods to generate dense pointmaps for 3D Gaussian prediction. For example, Splatt3R [63] builds upon the large-scale pretrained foundation 3D MASt3R model [4] by seamlessly integrating a Gaussian decoder, enabling pose-free feed-forward 3DGS. NoPoSplat [91] also utilizes the MASt3R as the backbone and predicts 3D Gaussians in a canonical space without ground-truth camera poses and depth. Large spatial model [92] combines DUSt3R [1] with a Gaussian prediction head and integrates additional semantic embeddings from the input images to enable feed-forward 3D Gaussian reconstruction. SmileSplat [93] utilizes the DUSt3R as the backbone to predict Gaussian surfels with a multi-head Gaussian regression decoder. SelfSplat [94] unifies DUSt3R-driven Gaussian prediction with self-supervised learning of depth and camera poses, enabling simultaneous prediction of geometry, pose, and Gaussian attributes. However, relying on DUSt3R and MASt3R imposes a limitation on these methods, as they inherit the constraint of pairwise inputs, restricting their scalability. PREF3R [95] builds on the pretrained reconstruction model Spann3R [55] for 3D Gaussian prediction and introduces a spatial memory network to achieve its functionality for multi-view images. However, the utilization of DUSt3R, MASt3R and Spann3R often leads to suboptimal rendering results due to imperfections in their geometry estimates.

Another research line of pointmap-based methods, FLARE [96], avoids using DUSt3R and MASt3R to obtain pointmaps, instead focusing on learning pointmaps for 3D Gaussian reconstruction and rendering. It still leverages pointmaps as the geometry representation and proposes the joint learning of camera poses, Gaussian pointmaps, enabling the high-quality feed-forward 3DGS reconstruction and rendering. Besides, LPGM [97] utilizes a pre-trained 3D diffusion model [20] to generate point clouds from a single-view input image, which are then processed by a dedicated point-to-Gaussian generator to produce the final 3D Gaussians.

2.4 Other 3D Representations

Except for the methods mentioned above, there have been several endeavors dedicated to the feed-forward reconstruction with different scene representations, exploring diverse research paths. In this section, we introduce several representative and advanced methods based on mesh, occupancy, and signed distance function (SDF) representations.

Mesh. As a widely used 3D representation, mesh has gained significant attention in feed-forward 3D reconstruction in recent years. For example, Pixel2Mesh [98] is proposed to produce 3D mesh from single input image, which leverages 2D CNN to extract image features for progressive mesh deformation. Mesh R-CNN [99] extends Mask R-CNN [100] by incorporating 3D shape inference. It introduces a voxel branch that predicts a coarse cubified mesh for each detected object, which is subsequently refined through a mesh refinement branch. Recently, one-2-3-45 [101] leverages the diffusion-based model Zero-1-to-3 [102] to produce multi-view images and feed these images to SDF-based generalizable neural surface reconstruction module [103] for feed-forward mesh reconstruction. One-2-3-45++ [104] enhances the consistency of synthesized multi-view images and utilizes a 3D diffusion-based module conditioned on multiple views to generate a textured mesh in a coarse-to-fine manner. However, they often suffer from low reconstruction quality with compromised geometry. To address this issue, Wonder3D [73] introduces a cross-domain diffusion model to generate multi-view-consistent normal maps and RGB images. By leveraging these consistent outputs, it reconstructs high-quality 3D meshes through a geometry fusion process. Unique3D [105] first employs a multi-view diffusion model alongside a normal diffusion model to generate multiview-consistent images and normal maps. It then introduces a fast and consistent mesh reconstruction algorithm that effectively integrates these outputs to produce high-quality 3D meshes with accurate geometry. Besides, several methods are proposed to utilize the strong capability of large reconstruction model [10] to achieve high-quality mesh reconstruction. For example, MeshLRM [106] integrates differentiable surface extraction and rendering into large reconstruction model, enabling the direct generation of high-fidelity 3D meshes from input images. InstantMesh [107] employs a multi-view diffusion model to synthesize novel views and utilizes a transformer-based large reconstruction model to generate a high-quality 3D mesh from the multi-view images. MeshFormer [11] leverages 3D voxel representations and combines 3D convolution with transformer-based LRM, leading to improved 3D mesh geometry by incorporating 3D-native designs.

Occupancy. Occupancy [109], [110] refers to the property that describes whether a given point in 3D space is inside or outside a surface or object. Several methods have been proposed to achieve feed-forward occupancy representation with generalization capabilities. For example, Any-Shot GIN [111] aims to model occupancy-based 3D implicit reconstruction with strong generalization capability. The method begins with front-back depth estimation to generate depth maps for constructing a voxel-based representation. Subsequently, it extracts 3D features from this volume to infer the occupancy of any 3D point in space. MCC [112] employs encoder-decoder architecture to reconstruct an occupancy-based representation. It first encodes a compressed representation of the scene appearance and geometry and then utilizes the encoded representation to predict occupancy probabilities and RGB colors for each 3D point. Additionally, Huang et al. introduce ZeroShape [113], a regression-based method for 3D occupancy reconstruction that achieves state-of-the-art performance in zero-shot generalization by intermediate geometric representation and explicit reasoning.

SDF. Signed Distance Function (SDF) [114] is a mathematical function that represents the geometry of a shape or surface in space. For any point in 3D space (or 2D), the SDF returns the shortest distance from that point to the surface of the object. The sign of the distance indicates whether the point is inside or outside the object. Several methods have been proposed to enable feed-forward SDF

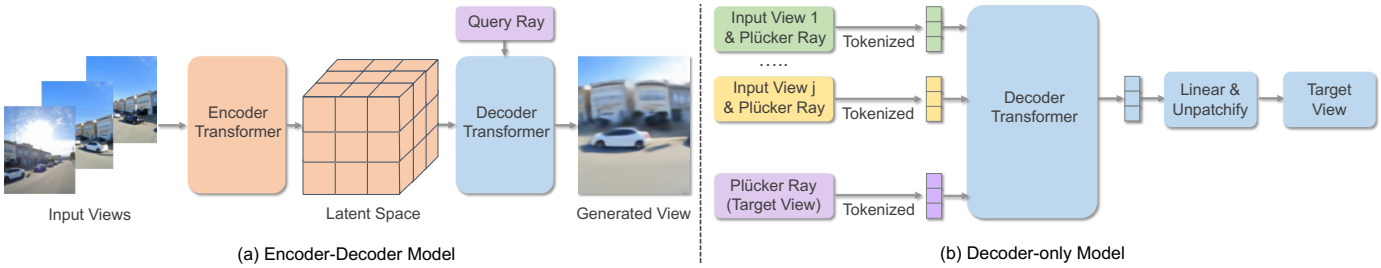

Fig. 5: Typical frameworks of regression-based representation-free models. The samples are adapted from [108] and [7].

representations. For example, Shap-E [20] transforms point clouds and RGBA input images into a sequence of latent vectors that serve as inputs for subsequent SDF prediction. SparseNeuS [103] initially builds a hierarchy of volumes that represent local surface details, which are then used to infer SDF-based surfaces through a progressive coarse-to-fine process. VolRecon [115] employs a view transformer to integrate features across multiple views and utilizes a ray transformer to estimate SDF values for points sampled along each ray. ReTR [116] also leverages transformer for SDF prediction. It instead introduces occlusion transformer and render transformer to fuse features and perform rendering. C2F2NeUS [117] incorporates multi-view stereo (MVS) into SDF-based surface reconstruction by first constructing a hierarchy of geometric frustums for each view to capture local-to-global scene geometry. Features extracted from these frustums are then fused using a cross-view and cross-level fusion strategy to facilitate accurate SDF prediction. UFORecon [118] also achieves MVS-based SDF reconstruction. It introduces cross-view matching transformer to extract cross-view matching features to construct hierarchical correlation volumes, enabling impressive SDF-based surface reconstruction under camera views with limited overlaps. To improve the reconstruction quality and training efficiency, CRM [119] incorporates geometric priors into network designs based on the spatial alignment between triplanes and the six input orthographic views. Specifically, it employs multi-view diffusion model to generate six synthesized orthographic images first and then introduces convolutional reconstruction model to map these views to triplane features, which are subsequently decoded into SDF values.

2.5 3D Representation-Free Models

Feed-forward representation-free models aim to directly feedforward synthesize novel views without 3D representations (e.g., NeRF and 3DGS). We broadly categorize the methods for two categories: regression-based methods (Sec. 2.5.1) and generative methods (Sec. 2.5.2).

2.5.1 Regression-based Feed-Forward View Synthesis

Regression-based feed-forward methods aim to formulate the rendering process as a regression problem, learning a rendering function (typically transformer-based neural network) to predict pixel colors of novel views from sparse-view inputs directly, without relying on 3D representations like NeRF or 3DGS. The key advantage of these methods is their ability to eliminate the inductive bias inherent in 3D representations. Based on their architecture, we classify these methods into two categories: encoder-decoder models and decoder-only models.

Encoder-Decoder Models. Scene representation transformer (SRT) [108], as a representative encoder-decoder model illustrated in Fig. 5(a), leverages a transformer-based encoder to map multi-view input images to latent representations first and then outputs novel-view images from a transformer-based decoder with light field rays. RUST [121] inherits an encoder-decoder architecture and enables novel view synthesis solely from RGB images, without the need for camera poses. Specifically, it queries the decoder using implicit latent poses predicted by a learned pose estimation module, rather than relying on explicit poses as in SRT. OSRT [122] focuses on object-centric 3D scenes and incorporates a slot attention module on SRT to map the encoded latent representations to object-centric slot representations where each slot corresponds to an object or a part of the background. Additionally, it replaces the SRT decoder with a slot mixer, enabling novel view rendering in a single forward pass, regardless of the number of slots. To extend SRT to large-scale scenes, RePAST [123] integrates relative camera pose information into the attention layer of SRT. However, these methods often suffer from degraded details and suboptimal rendering quality. To address this issue, several approaches incorporate geometric information to enhance model performance. For example, GPNR [124] integrates epipolar geometry within its encoder-decoder architecture, while Du et al. [125] introduce a multi-view vision transformer and epipolar line sampling to improve scene geometry. GBT [126] incorporates ray distance-based geometry reasoning into multihead attention layers of transformers in encoder and decoder. GTA [127] introduces geometric transform attention to embed the geometrical structure of tokens into the transformer and integrates it into SRT to enhance transformer-based rendering. However, despite the improved model performance, geometrical designs often integrate additional 3D inductive biases. LVSM [7] removes the geometrical designs and leverages transformer-based large reconstruction model with self-attention to enhance the capacity of encoder-decoder architecture and takes posed input images and Plücker ray embeddings to regress the target view pixels. To enable 3D view synthesis without any 3D supervision—such as ground-truth 3D geometry and camera poses—RayZer [128] is proposed. It adopts the encoder-decoder architecture of LVSM [7] and introduces a large, self-supervised multi-view 3D model that first learns camera parameters and latent scene representations from

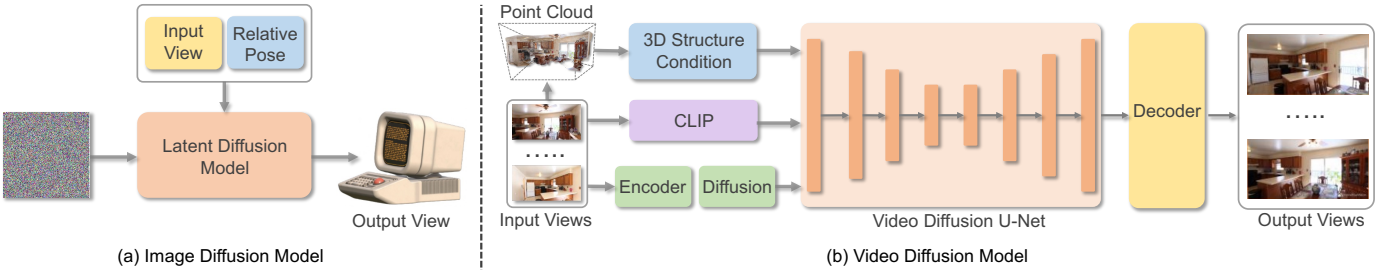

Fig. 6: Typical frameworks of generative representation-free models. The samples are adapted from [102] and [120].

unposed input images, and then renders novel views.

Decoder-Only Models. To minimize 3D inductive bias introduced by latent representations in encoder-decoder architectures, LVSM [7], as illustrated in Fig. 5(b), adopts a decoder-only design with a single-stream transformer, directly mapping input multi-view tokens to target view tokens.

2.5.2 Generative Feed-Forward View Synthesis

Regression-based methods work well for view interpolation, producing impressive visual results near the input views. However, it struggles with view extrapolation, leading to poor predictions for new views beyond the existing viewpoints, especially when estimating unseen regions of the scene. In contrast, generative feed-forward methods instead leverage generative models to synthesize realistic novel views based on learned data distributions, enabling view extrapolation even from a single input image. The generated multi-view images can often be utilized as dense inputs for high-fidelity NeRF- and 3DGS-based reconstruction.

Earlier works primarily leverage transformer-based autoregressive models [129], [130]. For example, GFVS [131] approaches novel view synthesis from single view by treating it as sampling target images from a learned distribution conditioned on a source image and camera transformation, where the distribution is modeled autoregressively by leveraging a VQGAN [130] with transformer. ViewFormer [132] extends single-view NVS of GFVS to multi-view NVS. Specifically, it first uses VQVAE codebook [129] to encode images into latent representations, then queries latent codes of target views and employs a transformer to map the latent codes to image tokens, which are subsequently decoded into novel views.

Recently, latent diffusion model [77] has been widely used in novel view synthesis due to its capability in generating high-resolution images, which encodes input images into a latent space by a pretrained variational autoencoder and applies diffusion within the latent representation and aims to learn the conditional distribution over the target image at the novel view. Diffusion-based generative methods have been widely explored using various diffusion priors, which will be introduced below.

Image Diffusion Model. Zero-1-to-3 [102] leverages the latent diffusion model [77] pretrained for text-to-image generation and replaces text embedding with relative camera poses as conditioning to achieve novel view synthesis as illustrated in Fig. 6(a). ZeroNVS [133] extends Zero-1-to-3 to achieve single-view scene-level novel view synthesis by finetuning it on diverse large-scale object and real-scene datasets [23], [134], [135]. To enhance pose conditioning accuracy, ReconFusion [136] integrates PixelNeRF and modifies the pre-trained latent diffusion model to utilize PixelNeRF-rendered features derived from relative camera poses as conditioning. However, these methods still face challenges in generating consistent novel views. To address this limitation, SyncDreamer [137] initializes the diffusion model with pretrained Zero-1-to-3 weights and extends the diffusion model to capture the joint probability distribution of multi-view images, generating novel views with multi-view consistency. Zero123++ [12] arranges six surrounding views into a single image, facilitating accurate joint distribution modeling of an object's multi-view representations. Consistent123 [138] combines Zero-1-to-3 and stable diffusion to provide diffusion priors to ensure the multiview-consistent synthesized novel views. ConsistNet [139] builds on Zero-1-to-3 as its backbone and performs multiple diffusions in parallel, each handling a specific viewpoint. To enforce multi-view geometric consistency, it incorporates a dedicated plug-in block that aligns the generated images accordingly. MVDream [74] proposes a multiview diffusion model that leverages both 2D and 3D data, combining the generalizability of 2D diffusion models with the consistency of 3D renderings. It further demonstrates that a multiview diffusion model can implicitly provide generalizable 3D priors without specific 3D representations.

Video Diffusion Model. Video diffusion models [140], [141] have achieved impressive realism in video synthesis and are thought to inherently and implicitly capture 3D structures. Building on this capability, recent approaches have explored leveraging their priors to generate multi-view images for high-quality 3D reconstruction. For example, ReconX [120], as illustrated in Fig. 6(b), harnesses the powerful generative prior of large pretrained video diffusion models [141] for generalizable sparse-view reconstruction. It encodes extracted point clouds as 3D structural conditions, ensuring multi-view consistency in the generated novel views. Similarly, ViewCrafter [8] first builds a point cloud representation and then utilizes point cloud renderings as the conditions of video diffusion model [141] to enable consistent and accurate novel view synthesis. MultiDiff [142] leverages a single reference image and a predefined target camera trajectory as conditions, utilizing depth cues to encourage consistent novel view synthesis. More recently, GenFusion [143], DifFusion3D+ [144], and SpatialCrafter [145] have bridged reconstruction and generation by using video diffusion models as scene recon-

struction refiners, enabling both artifact removal and scene content expansion.

3 Tasks & Applications

3.1 3D-aware Image Synthesis

3D-aware image synthesis refers to generating 2D images guided by an understanding of underlying 3D geometry, enabling the image synthesis of objects or scenes from different viewpoints while maintaining 3D consistency. These methods typically adopt a GAN-based framework, where a 3D representation is reconstructed in a feed-forward manner and then rendered to synthesize realistic 2D images.

Several earlier methods employ voxel-based representations (e.g., PlatonicGAN [146]) or 3D feature representations (e.g., HoloGAN [147] and BlockGAN [148]). However, these approaches often suffer from limited multi-view consistency. To address this, GRAF [149] introduces a generative radiance field as a 3D scene representation, significantly improving consistency across different viewpoints. It designs a conditional NeRF that takes shape and appearance latent codes as conditions and produces images via volume rendering. PiGAN [150] leverages implicit neural representations with periodic activation functions to model scenes as view-consistent radiance fields. Subsequently, GIRAFFE [151] constructs compositional generative radiance fields for scene representations, enabling controllable image synthesis. StyleNeRF [152] combines NeRF-based 3D scene representations with a style-based generative model for high-resolution, 3D-consistent image synthesis. EG3D [153] introduces a explicit-implicit triplane representation to achieve efficient and high-quality 3D-aware image synthesis.

Due to the high computational cost of volume rendering in implicit NeRF-based scene representations, Hyun et al. propose GSGAN [154], which replaces NeRF with 3D Gaussian Splatting (3DGS), enabling more efficient scene rendering through rasterization-based splatting. To stabilize the training of 3DGS-based 3D-aware image synthesis, GSGAN introduces hierarchical Gaussian representations, enabling coarse-to-fine scene modeling.

3.2 Camera-controled Video Generation

To enable camera pose control in the video generation process, MotionCtrl, CameraCtrl, I2VControl-Camera inject the camera parameters (extrinsic, Plücker embedding, or point trajectory) into a pretrained video diffusion model. Building upon this, CamCo integrates epipolar constraints into attention layers, while CamTrol, NVS-Solver, and ViewExtrapolator leverage explicit 3D point cloud renderings to guide the sampling process of the video diffusion models in a training-free manner. AC3D carefully design the camera representation injection to the pretrained model. ViewCrafter and Gen3C fine-tuned video diffusion models on point cloud renderings to enable better novel view synthesis. VD3D enables camera control to transformer-based video diffusion models. Beyond static scenes, CameraCtrl II, and ReCamMaster enable camera-controled video generation on dynamic scenes by conditioning the video diffusion models on camera extrinsic parameters, while TrajectoryCrafter also enables dynamic scene view synthesis by conditioning the video diffusion models on dynamic point cloud. Several recent works have advanced beyond single-camera scenarios: CVD, Caiva, Vivid-ZOO, and SynCamMaster have developed frameworks for multi-camera synchronization.

3.3 Pose-free 3D Reconstruction

The development of feed-forward models has enabled the reconstruction of 3D scenes from unposed images or videos without the need for per-scene optimization. FlowCam [155] employs a single-view feed-forward generalizable NeRF to generate point maps for different input viewpoints, using optical flow to estimate poses and integrate point maps from multiple views to reconstruct neural radiance fields. CoPoNeRF [156] performs pair matching at the feature map level, extracting multi-level features from image pairs to construct 4D correlation maps encoding pixel-pair similarities. These maps are further refined for flow and pose estimation, enabling the renderer to compute color and depth from the refined features and estimated poses.

To extend these approaches into 3D Gaussian Splatting (3DGS), GGRt [78] employs PixelSplat [5] for predicting viewpoint-specific 3D Gaussian maps and introduces a pose estimation module that jointly optimizes camera poses alongside Gaussian predictions. PF3plat [87] proposes a coarse-to-fine strategy, estimating depth, confidence, and camera poses from input images to guide the prediction of 3D Gaussians using multi-stereo and guidance cost volumes.

Additionally, several methods build upon DUSt3R [1] for pose-free 3D reconstruction. DUSt3R itself, as a pioneering feed-forward method, utilizes a transformer-based architecture to regress 3D point maps directly from image pairs. Spann3R [55] augments DUSt3R with a spatial memory network, allowing multi-view inputs and improving efficiency by eliminating global alignment. However, Spann3R's sequential processing introduces error accumulation in reconstruction. Fast3R [53] overcomes this limitation by introducing a global fusion transformer, processing multiple views simultaneously and significantly enhancing reconstruction quality. Conversely, CUT3R [3] refines sequential reconstruction by maintaining and incrementally updating a persistent internal state that encodes scene content. Instead of relying on pairwise feature matching with previous views, CUT3R updates its internal state continuously and utilizes it directly to predict the pointmap of the current view.

Moving beyond pointmap-level reconstruction, several methods have further developed high-quality novel view synthesis through 3D Gaussian reconstruction. Splatt3R [63] extends DUSt3R by adding a Gaussian head decoder that predicts Gaussian parameters directly from image pairs. LSM [92] similarly integrates a Gaussian head and further incorporates semantic embeddings from input images to augment anisotropic Gaussian predictions. NoPosplat [91], after integrating a Gaussian head, performs full-parameter training to predict 3D Gaussians in a canonical space without relying on ground-truth camera poses or depth. PREF3R [95], based on Spann3R, also adds a Gaussian head to achieve multi-view input 3D Gaussian predictions.

SmileSplat [93], another Spann3R derivative, opts to predict Gaussian surfels instead of traditional 3D Gaussians. SelfSplat [94] integrates DUSt3R-based Gaussian predictions with self-supervised depth and pose estimation, jointly predicting depth, camera poses, and Gaussian attributes in a unified neural network. Lastly, FLARE [96] incorporates additional modules for pose estimation and global geometry projection, facilitating alignment of DUSt3R-based network token outputs.

3.4 Dynamic 3D Reconstruction

Compared to static scene reconstruction, dynamic scene reconstruction poses significant challenges mainly due to the presence of moving objects, changing viewpoints, and temporal variations in scene geometry. Extending feed-forward 3D reconstruction for dynamic scenarios mainly involves robust pose estimation to mitigate moving object interference, together with dynamic area segmentation for updating changing environments.

Seminal work on monocular depth estimation methods learned to predict temporal consistent depth video using temporal attention layers [157] and generative priors [158], [159]. Though they demonstrate pleasure 3D points on camera space, however, they fails to provide global scene geometry due to the lack of camera pose estimation.

To jointly resolve pose and obtain point cloud at canonical space, Robust-CVD [160] and CasualSAM [161] integrates a depth prior with geometric optimization to estimate a smooth camera trajectory, as well as detailed and stable depth and motion map reconstruction. Most recently, MegaSaM [162] further improves pose and depth accuracy by combining the strengths of several prior works including DROID-SLAM [163], optical flow [164] and monocular depth estimation model [165], leading to results with previously unachievable quality.

Alternatively, instead of taking advantage from monocular prior models, some methods aim to train a dynamic 3D model from multi-view 3D reconstruction models, *e.g.*, DUSt3R [1]. MonST3R [166] estimates pointmap at each timestep and processes them using a temporal sliding window to compute pairwise pointmap for each frame pair with MonST3R and optical flow from off-the-shelf method. These intermediates then serve as inputs to optimize a global point cloud and per-frame camera poses and intrinsics. Video depth can be directly derived from this unified representation. To speed up the optimization process in MonST3R, DAS3R [167] trains a dense prediction transformer [168] for motion segmentation inference and model the static scene as Gaussian splats with dynamics-aware optimization, allowing for more accurate background reconstruction results. Recent work CUT3R [3] fine-tunes MonST3R [166] on both static and dynamic datasets, achieving feedforward reconstruction but without predicting dynamic object segmentation, thereby entangling the static scene with dynamic objects. Although effective, these methods require costly training on diverse motion patterns to generalize well. In contrast, Easi3R [169] takes an opposite path, exploring a training-free and plug-in-play adaptation that enhances the generalization of DUSt3R variants for dynamic scene reconstruction, achieving accurate dynamic region segmentation, camera pose estimation, and 4D dense point map reconstruction at almost no additional cost on top of DUSt3R.

Another line of research focuses on leveraging video pre-trained models for point map prediction by modeling 3D scenes as geometry videos. These approaches utilize diffusion models to learn the joint distribution of multi-view RGB and geometric frames. A geometry video consists of standard RGB channels augmented with geometry channels, which encode structural information such as depth [170], XYZ coordinates [171], color point rendering [172], [173], or a combination of point-depth-ray maps [174]. Notably, Aether [175] presents a unified framework that takes as input both image and action latents — such as ray maps — and produces predictions for images, actions, and depth. By flexibly combining different input conditions, Aether successfully achieved 4D dynamic reconstruction from video-only input, image-to-video generation from a single image, and camera-conditioned video synthesis given an image and a camera trajectory.

To enable 3D point tracking, Stereo4D [176] proposes a dynaDUSt3R architecture by incorporating a motion head for scene flow prediction. They use stereo videos from the internet to create a dataset of over 100,000 real-world 4D scenes with metric scale and long-term 3D motion trajectories for training. Instead of predicting point map and flow map at reference and target viewpoints, St4RTrack [177] outputs two point maps of different time steps for the reference view given two dynamic frames. The network is trained by reprojected supervision signals including 2D trajectories and monocular depth, without the need for direct scene flow annotation.

3.5 3D Understanding

There have been works that embed features (e.g., language features obtained via CLIP) into feed-forward 3D reconstruction models, enabling 3D querying and segmentation through feature representations. Among earlier efforts, Large Spatial Model [178] employs a point-based transformer that facilitates local context aggregation and hierarchical fusion to reconstruct a set of semantic, anisotropic 3D Gaussians in a supervised, end-to-end manner. GSemSplat [179] introduces a semantic head that predicts both region-specific and context-aware semantic features, which are then decoded into high-dimensional representations using MLP blocks for open-vocabulary semantic understanding. In contrast to these two works, which focus on open-vocabulary segmentation, SplatTalk [180] tackles the broader challenge of free-form language reasoning required for 3D visual question answering (3D-VQA). It incorporates a feed-forward feature field as a submodule, including training a Gaussian encoder and a Gaussian latent decoder to reconstruct a 3D-language Gaussian field.

3.6 Matching & Optical Flow

Recent advances in feed-forward 3D reconstruction have led to significant progress in image matching. One notable example is MASt3R, which builds upon the DUSt3R foundation model to enable efficient and robust image matching

in a single forward pass. By augmenting the DUSt3R architecture with a dedicated head for dense local feature extraction, MASt3R introduces a mechanism to improve matching accuracy while maintaining the robustness characteristic of pointmap-based regression. To mitigate the computational cost, MASt3R further incorporates a novel reciprocal matching scheme that reduces the quadratic complexity typical of dense matching.

On the other hand, MASt3R is fundamentally limited to processing image pairs with poor scalability for large image collections. To handle this issue, MASt3R-SfM proposes to leverage the frozen encoder of MASt3R for image retrieval, enabling it to process large and unconstrained image collections with quasi-linear complexity scalablely. Importantly, the robustness of MASt3R's local reconstructions allows the SfM pipeline to dispense with traditional RANSAC-based filtering. Instead, optimization is performed through successive gradient-based refinement in both 3D space (via a matching loss) and 2D image space (via reprojection loss), thus highlighting the potential of feed-forward paradigms to serve as both matching engines and geometric optimizers.

3.7 Digital Human

Recent progresses in feed-forward 3D reconstruction have attracted increasing attention in photo-realistic 3D avatar. For example, GPS-Gaussian [181] defines 2D Gaussian parameter maps on the input views and directly predicts 3D Gaussians in a feed-forward manner, enabling efficient and generalizable human novel view synthesis. Avat3r [182] builds upon the Large Gaussian Reconstruction Model [66] to predict 3D Gaussians corresponding each pixel of input image, achieving animatable 3D reconstruction and high-quality 3D head avatars. Besides, Avat3r also incorporates priors from DUSt3R [1] and the human foundation model Sapiens [183] to further enhance generalization and robustness in 3D head avatar reconstruction.

3.8 SLAM

Recent SLAM systems have increasingly adopted feed-forward models to replace traditional geometric pipelines, offering real-time and dense reconstruction from monocular RGB videos. MASt3R-SLAM [184] leverages the MASt3R [4] prior to build a real-time dense monocular SLAM system that operates without requiring known camera calibration. It integrates efficient techniques for pointmap matching, Sim(3)-based camera tracking, local pointmap fusion, loop closure, and second-order global optimization to maintain pose and map consistency. Similarly, based on DUSt3R, SLAM3R [54] introduces a real-time, end-to-end dense reconstruction system that directly predicts 3D pointmaps from RGB videos. Its Image-to-Points (I2P) module extends DUSt3R to multi-view inputs for improved local geometry, while the Local-to-World (L2W) module incrementally aligns local pointmaps into a global frame—eliminating the need for camera pose estimation or global optimization. However, MASt3R and DUSt3R, being inherently two-view, limits each inference to a fixed image pair, making large-scale fusion dependent on iterative matching and optimization. VGGT-SLAM [185] addresses this limitation by adopting the more powerful VGGT transformer, which supports arbitrary-length image sets (within memory constraints) and jointly predicts dense point clouds, camera poses, and intrinsics in a single forward pass. This allows VGGT-SLAM to construct larger submaps and align them via projective transformations optimized on the SL(4) manifold.

3.9 Robotics

ManiGaussian [186] adopts a feed-forward model for robotics manipulation. It introduces a dynamic GS framework to model the propagation of diverse semantic features, along with a Gaussian world model that supervises learning by reconstructing future scenes for scene-level dynamics mining. Its follow-up work ManiGaussian++ [187], extends ManiGaussian by introducing the hierarchical Gaussian world model to learn the multibody spatiotemporal dynamics for bimanual tasks. While many works use optimization-based NeRF and 3D Gaussians for robotics tasks like manipulation and navigation, few adopt feed-forward 3D models due to reconstruction quality concerns. However, as feed-forward reconstruction quality rapidly improves, more works are expected to shift toward these models for their significantly faster inference speed.

4 Experiment

4.1 Datasets

Datasets are the core of feed-forward 3D reconstruction and view synthesis. To give an overall picture of the datasets, we tabulate the detailed scene and annotation types in popular datasets in Table 1. The scene type are divided into objects, indoor scenes and outdoor scenes. And we also indicate synthetic datasets (e.g., ShapeNet [188], Objaverse [189] and Virtual KITTI2 [190]), real-world datasets (e.g., DL3DV-10K [191], ACID [135] and RealEstate10K [134]), static datasets (e.g., ScanNet [192], MVImgNet [193] and ARKitScenes [194]) and dynamic datasets (e.g., Waymo [195], KITTI360 [196] and PointOdyssey [197]). Notably, several datasets, such as DL3DV-10K [191] and TartanAir [198], include both static and dynamic scenes.

4.2 Evaluation Metrics

To achieve a faithful evaluation, comprehensive metrics are used to evaluate different downstream tasks in feed-forward 3D vision.

For novel view synthesis evaluation, SSIM (Structural Similarity Index), PSNR (Peak Signal-to-Noise Ratio), and LPIPS (Learned Perceptual Image Patch Similarity) [217] are commonly used to evaluate image quality, but they focus on different perspectives. PSNR converts the MSE between the rendered view and its ground truth reference into a decibel scale, which calculates pixel-wise similarity. SSIM measures perceptual similarity using three local components, luminance, contrast, and structure, which provides patch-level similarity. LPIPS estimates perceptual distance in a learned feature space by computing the L2 difference between deep features of the two images in a pretrained network, which measures feature-level similarity.

For camera pose estimation, RTA (Relative Translation Accuracy), RRA (Relative Rotation Accuracy), and AUC (Area Under Curve) are usually reported. RTA and RRA

TABLE 1: Summarization of popular datasets for feed-forward 3D reconstruction and view synthesis.

Datasets	#Scenes (Objects)	Type	Real	Static	Dynamic	Camera	Point Cloud	Depth	Mesh	LiDAR	Semantic	Mask	Optical Flow
DTU [199]	124	Objects	Real	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
Pix3D [200]	395	Objects	Real	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
GSO [201]	1,030	Objects	Real	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗
OmniObject3D [202]	6,000	Objects	Synthetic	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗
CO3D [23]	18,619	Objects	Real	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗
WildRGBD [203]	23,049	Objects	Real	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗
ShapeNet [188]	51,300	Objects	Synthetic	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗
MVImgNet [193]	219,188	Objects	Real	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗
Objaverse [189]	818K	Objects	Synthetic	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗
Objaverse-XL [204]	10.2M	Objects	Synthetic	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗
7Scenes [205]	7	Indoor Scenes	Real	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗
Replica [206]	18	Indoor Scenes	Real	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗
TUM RGBD [207]	39	Indoor Scenes	Real	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
Matterport3D [208]	90	Indoor Scenes	Real	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗
HyperSim [209]	461	Indoor Scenes	Synthetic	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗
Dynamic Replica [210]	524	Indoor Scenes	Synthetic	✗	✓	✓	✗	✓	✗	✗	✗	✓	✓
ScanNet++ [211]	1,006	Indoor Scenes	Real	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗
ScanNet [192]	1,513	Indoor Scenes	Real	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗
ARKitScenes [194]	1,661	Indoor Scenes	Real	✓	✗	✓	✓	✓	✓	✓	✗	✗	✗
Virtual KITTI2 [190]	5	Outdoor Scenes	Synthetic	✗	✓	✓	✗	✓	✗	✗	✓	✗	✓
KITTI360 [196]	11	Outdoor Scenes	Real	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗
Spring [212]	47	Outdoor Scenes	Synthetic	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓
MegaDepth [213]	196	Outdoor Scenes	Real	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
ACID [135]	13,047	Outdoor Scenes	Real	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
MipNeRF360 [214]	9	Indoor and Outdoor Scenes	Real	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
Tanks&Temples [215]	21	Indoor and Outdoor Scenes	Real	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
ETH3D [216]	25	Indoor and Outdoor Scenes	Real	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗
PointOdyssey [197]	159	Indoor and Outdoor Scenes	Synthetic	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗
TartanAir [198]	1,037	Indoor and Outdoor Scenes	Synthetic	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓
DL3DV-10K [191]	10,510	Indoor and Outdoor Scenes	Real	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
RealEstate10K [134]	74,766	Indoor and Outdoor Scenes	Real	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
BlendedMVS [189]	113	Objects, Indoor and Outdoor Scenes	Synthetic	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗

measures the relative angular errors in translation and rotation between two image pairs respectively. AUC computes the area under the accuracy curve across different angular thresholds, where accuracy at each threshold is given by the proportion of predicted pose whose angular error relative to the ground truth pose falls below that threshold.

For monocular depth estimation, people usually calculate the absolute relative error $|y - \hat{y}|/y$ where y is ground-truth and $\hat{y}$ the prediction, and percentage of inlier points $\delta < 1.25$, which is the percentage of predicted depths within a 1.25-factor of true depth.

In point map evaluation, the standard metrics include pointcloud Accuracy (or precision), Completeness (or recall), and Chamfer distance. Accuracy is the average nearest-neighbor distance from each predicted point to the ground-truth surface, indicating how precisely predicted points are placed; Completeness is the average nearest-neighbor distance from each ground-truth point to the reconstruction, reflecting how fully the ground-truth surface is covered. Chamfer Distance combines the Accuracy and Completeness score and is thus more comprehensive. The three metrics can also be adapted to evaluate multiview depth prediction in a similar way.

For dynamic point tracking, OA (Occlusion Accuracy), σ^{vis}_{avg}, and AJ (Average Jaccard) are used together. OA measures the binary accuracy of occlusion predictions; σ^{x}_{avg} measures the fraction of points that are accurately tracked within a certain pixel threshold; Average Jaccard considers both occlusion and prediction accuracy.

5 Open Challenges

Though Feed-forward 3D models has made notable progress and achieved superior performance in recent years, there exist several challenges for future exploration. In this section, we overview the typical challenges, share our humble opinions on possible solutions, and highlight the future research directions.

5.1 Limited Modality in Datasets

A critical limitation of current 3D reconstruction and view synthesis datasets lies in their incomplete modality coverage. Many widely-used benchmarks, such as RealEstate10K [134] and MipNeRF360 [214], provide only RGB camera images, while omitting essential complementary signals like depth, LiDAR, or semantic annotations. Even large-scale collections like Objaverse-XL [204] (10.2M objects) focus primarily on synthetic mesh data, lacking

the real-world sensor modalities needed to train robust models. This imbalance forces researchers to combine disparate datasets, which inevitably introduces domain shifts and annotation inconsistencies. The modality gap is particularly acute for dynamic scene understanding. While several datasets provide dynamic sequences, those with comprehensive multi-modal annotations (e.g., synchronized RGB, depth, optical flow, and 3D motion) remain significantly fewer than their static counterparts. Most dynamic datasets prioritize either camera motion or object movement, but rarely capture both simultaneously with full sensor suites. This scarcity of richly-annotated dynamic data severely constrains the development of models capable of handling real-world scenarios where both cameras and objects move freely. A fundamental challenge emerges: how to create scalable, modality-rich datasets that combine the diversity of synthetic collections like Objaverse [189] with the multi-sensor completeness of real-world benchmarks such as ScanNet++ [211] or Waymo [195]. Current approaches must rely on patching together incompatible data sources, which ultimately limits progress toward generalizable 3D understanding. The field urgently needs comprehensive resources that provide aligned multi-modal signals, including RGB, depth, semantics, and motion, under unified collection protocols to overcome these limitations.

5.2 Reconstruction Accuracy

Feed-forward 3D reconstruction models have made noticeable progress over recent years. However, their reconstruction accuracy—particularly in terms of depth map precision—is still inferior to traditional multi-view stereo (MVS) methods [29], [218], [219], which explicitly utilize camera parameters for all input frames. Specifically, MVS approaches typically leverage known camera parameters and hypothesized depth sets to construct cost volumes, subsequently processed to predict accurate depth or disparity maps. An intriguing hypothesis is, feed-forward 3D reconstruction models might spontaneously learn an approximation of such cost volumes. Modern feed-forward reconstruction models [2], [53] mostly employ self-attention layers, theoretically enabling them to approximate or even exceed the representational capacity of traditional cost volumes. With sufficient high-quality training data, these feed-forward models have the potential to match or surpass the accuracy of MVS-based methods. Moreover, incorporating explicit camera parameters or additional priors into the feature backbone, such as through Diffusion Transformers (DiT) [220], offers another promising avenue to enhance reconstruction accuracy. Consequently, we anticipate that feed-forward models will continue to evolve, eventually much outperforming traditional MVS methods and achieving sensor-level accuracy, comparable to technologies like LiDAR or high-precision scanning systems.

5.3 Free-viewpoint Rendering

The free-viewpoint synthesis challenge in 3D reconstruction lies in the difficulty of generating high-quality novel views far from observed viewpoints, primarily due to disocclusions, geometric uncertainty, and limited generalization of feed-forward models. When extrapolating beyond the input camera distribution, unseen regions often lead to artifacts such as blurring, ghosting, or incorrect geometry, as existing methods rely heavily on local consistency and struggle to infer plausible content for occluded areas. Additionally, view-dependent effects and complex light transport further complicate synthesis, requiring models to reason beyond interpolation-based priors. Addressing these challenges demands advancements in scene understanding, robust geometric priors, and techniques that can hallucinate missing details while maintaining consistency across novel viewpoints.

5.4 Long Context Input

Existing methods often rely on full attention layers for 3D geometry reasoning and novel view synthesis, resulting in a cubic increase in token count and computing cost. For instance, inferring from 32 images using VGGT consumes approximately xxG FLOPS, while inferring 100 images can cost xxG FLOPS. In practice, training the models on larger than 32 views still difficulties fitting in even the most advanced GPU devices. A promising alternative is the use of recurrent mechanisms, such as in Cut3R, which incrementally incorporate views while maintaining a state memory. Although this approach significantly reduces memory consumption, it suffers from the problem of forgetting previously seen information. Efficiently reasoning over hundreds or even thousands of views in a memory- and computation-efficient manner remains an open challenge.

6 Social Impacts

Feed-forward 3D reconstruction and view synethesis has gained considerable attention in recent years due to their wide-ranging applications across various industries. In this section, we will discuss its applications and misuses from a societal aspect.

6.1 Applications

3D reconstruction models are creating significant positive societal impacts. To name a few, they have the potential to transform the film and gaming industries with more realistic visual effect and production speed by using reconstructed or generated 3D assets. They are also valuable in the development of smart cities, where they can be used to create ”digital twins” of critical infrastructure for simulation and maintenance planning. Additionally, 3D reconstruction can help cultural heritage preservation, as they allow ancient artifacts and statues to be digitally preserved before they deteriorate.

6.2 Misuse

The widespread availability of 3D reconstruction models raises significant privacy concerns. For example, private property could be reconstructed without the owner’s permission simply by taking a few pictures. To address these issues, new regulations should be established as 3D reconstruction technologies become increasingly accessible. In addition, the generative capabilities of feed-forward 3D reconstruction models can be misused to create false evidence,

such as fabricated crime scenes. To prevent such misuse, research should focus on developing detection models that can distinguish between generated and real content. People can also develop techniques to add "invisible watermarks" on generated outputs, enabling simple decoding to verify if content is artificially created.

6.3 Environment

Feed-forward 3D reconstruction models inherently demand substantial GPU resources and energy because they usually need to learn generic scene priors from large-scale datasets. Their inference stage, though, is more efficient: unlike optimization-based methods that update network weights at runtime, feed-forward models produce the results in a single pass within seconds. To further reduce computational costs, a general research direction is improving model generalizability. In this way, a pretrained model that generalizes well across different datasets can greatly speed up model training for downstream applications by offering rich semantic information.

7 Conclusion

Feed-forward 3D reconstruction and view synthesis have redefined the landscape of 3D vision, enabling real-time, generalizable, and scalable 3D understanding across a wide range of tasks and applications. This review covers the main approaches in feed-forward 3D reconstruction and view synthesis. Specifically, we provide an overview of these methods based on their underlying representations, such as NeRF, 3DGS, and PointMap. We also compare these methods by analyzing their strengths and weaknesses, aiming to inspire new paradigms that leverage the advantages of existing frameworks. In addition, we discuss the tasks and applications of the feed-forward approaches, ranging from image and video generation to various types of 3D reconstruction. We also introduce commonly used datasets and evaluation metrics for assessing the performance of 3D feed-forward models in these tasks. Finally, we summarize the open challenges and future directions, including the need for more diverse modalities, free-viewpoint synthesis, and long-context generation.

Acknowledgments

Jiahui Zhang, Muyu Xu, Kunhao Liu and Shijian Lu

References

[1] S. Wang et al. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[2] J. Wang et al. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[3] Q. Wang et al. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[4] V. Leroy et al. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

[5] D. Charatan et al. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

[6] H. Liang et al. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. *arXiv preprint arXiv:2412.03526*, 2024.

[7] H. Jin et al. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] W. Yu et al. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[9] S. Szymanowicz et al. Bolt3d: Generating 3d scenes in seconds. *arXiv preprint arXiv:2503.14445*, 2025.

[10] Y. Hong et al. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

[11] M. Liu et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[12] R. Shi et al. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

[13] B. Mildenhall et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[14] B. Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[15] W. Jang and L. Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12949–12958, 2021.

[16] A. Trevithick and B. Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15182–15192, 2021.

[17] A. Chen et al. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14124–14133, 2021.

[18] A. Yu et al. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

[19] K. Rematas et al. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021.

[20] H. Jun and A. Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[21] A. Nichol et al. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[22] Q. Wang et al. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2021.

[23] J. Reizenstein et al. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

[24] J. Chibane et al. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7911–7920, 2021.

[25] P. Wang et al. Is attention all that nerf needs? *arXiv preprint arXiv:2207.13298*, 2022.

[26] Y. Chen et al. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023.

[27] H. Yang et al. Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16508–16517, 2023.

[28] T. Chen et al. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

[29] Y. Yao et al. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

[30] X. Gu et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2020.

[31] S. Cheng et al. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2524–2534, 2020.

[32] M. M. Johari et al. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18365–18375, 2022.

[33] Y. Liu et al. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7824–7833, 2022.

[34] M. Xu et al. Wavenerf: Wavelet-based generalizable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18195–18204, 2023.

[35] H. Lin et al. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

[36] H. Xu et al. Murf: multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20041–20050, 2024.

[37] T. Liu et al. Geometry-aware reconstruction and fusion-refined rendering for generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7654–7663, 2024.

[38] P. Wang et al. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023.

[39] D. Tochilkin et al. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

[40] D. Xie et al. Lrm-zero: Training large reconstruction models with synthesized data. *arXiv preprint arXiv:2406.09371*, 2024.

[41] J. Li et al. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

[42] D. Podell et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[43] Y. Xu et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023.

[44] T. Anciukevičius et al. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12608–12618, 2023.

[45] K.-E. Lin et al. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 806–815, 2023.

[46] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[47] J. Li et al. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12578–12588, 2021.

[48] R. Tucker and N. Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 551–560, 2020.

[49] E. Brachmann et al. Dsac-differentiable ransac for camera localization. In *CVPR*, pp. 6684–6692, 2017.

[50] E. Brachmann and C. Rother. Learning less is more-6d camera localization via 3d surface regression. In *CVPR*, pp. 4654–4662, 2018.

[51] E. Brachmann and C. Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021.

[52] S. Dong et al. Visual localization via few-shot scene region classification. In *2022 International Conference on 3D Vision (3DV)*, pp. 393–402. IEEE, 2022.

[53] J. Yang et al. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *arXiv preprint arXiv:2501.13928*, 2025.

[54] Y. Liu et al. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16651–16662, 2025.

[55] H. Wang and L. Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.

[56] W. Jang et al. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1071–1081, 2025.

[57] S. Li et al. Rig3r: Rig-aware conditioning for learned 3d reconstruction. *arXiv preprint arXiv:2506.02265*, 2025.

[58] S. Elflein et al. Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16774–16784, 2025.

[59] S. Liu et al. Regist3r: Incremental registration with stereo foundation model. *arXiv preprint arXiv:2504.12356*, 2025.

[60] S. Szymanowicz et al. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10208–10217, 2024.

[61] A. Chen et al. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*. Springer, 2024.

[62] Z.-X. Zou et al. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10324–10335, 2024.

[63] B. Smart et al. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

[64] O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

[65] S. Szymanowicz et al. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024.

[66] Y. Xu et al. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024.

[67] K. Zhang et al. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

[68] C. Ziwen et al. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint arXiv:2410.12781*, 2024.

[69] T. Dao and A. Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

[70] J. Xu et al. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. *arXiv preprint arXiv:2412.09573*, 2024.

[71] Z. Min et al. Epipolar-free 3d gaussian splatting for generalizable novel view synthesis. *arXiv preprint arXiv:2410.22817*, 2024.

[72] J. Tang et al. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

[73] X. Long et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

[74] Y. Shi et al. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[75] P. Wang and Y. Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

[76] C. Wewer et al. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2024.

[77] R. Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[78] H. Li et al. Ggrt: Towards generalizable 3d gaussians without pose priors in real-time. *arXiv e-prints*, pp. arXiv–2403, 2024.

[79] Y. Chen et al. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024.

[80] P. Pham et al. Mvgaussian: High-fidelity text-to-3d content generation with multi-view guidance and surface densification. *arXiv preprint arXiv:2409.06620*, 2024.

[81] C. Zhang et al. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9869–9877, 2025.

[82] H. Xu et al. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024.

[83] S. Tang et al. Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. *arXiv preprint arXiv:2410.06245*, 2024.

[84] C. Zhang et al. Pansplat: 4k panorama synthesis with feed-forward gaussian splatting. *arXiv preprint arXiv:2412.12096*, 2024.

[85] Y. Chen et al. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *arXiv preprint arXiv:2411.04924*, 2024.

[86] A. Blattmann et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[87] S. Hong et al. Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024.

[88] T. Lu et al. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *CVPR*, 2024.

[89] B. Zhang et al. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *arXiv preprint arXiv:2403.19655*, 2024.

[90] D. Xu et al. Agg: Amortized generative 3d gaussians for single image to 3d. *arXiv preprint arXiv:2401.04099*, 2024.

[91] B. Ye et al. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

[92] Z. Fan et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems*, 37:40212–40229, 2024.

[93] Y. Li et al. Smilesplat: Generalizable gaussian splats for unconstrained sparse images. *arXiv preprint arXiv:2411.18072*, 2024.

[94] G. Kang et al. Selfsplat: Pose-free and 3d prior-free generalizable 3d gaussian splatting. *arXiv preprint arXiv:2411.17190*, 2024.

[95] Z. Chen et al. Pref3r: Pose-free feed-forward 3d gaussian splatting from variable-length image sequence. *arXiv preprint arXiv:2411.16877*, 2024.

[96] S. Zhang et al. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. *arXiv preprint arXiv:2502.12138*, 2025.

[97] L. Lu et al. Large point-to-gaussian model for image-to-3d generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10843–10852, 2024.

[98] N. Wang et al. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67, 2018.

[99] G. Gkioxari et al. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9785–9795, 2019.

[100] K. He et al. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[101] M. Liu et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023.

[102] R. Liu et al. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

[103] X. Long et al. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pp. 210–227. Springer, 2022.

[104] M. Liu et al. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10072–10083, 2024.

[105] K. Wu et al. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[106] X. Wei et al. Meshlrm: Large reconstruction model for high-quality meshes. *arXiv preprint arXiv:2404.12385*, 2024.

[107] J. Xu et al. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[108] M. S. Sajjadi et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6229–6238, 2022.

[109] L. Mescheder et al. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

[110] S. Peng et al. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 523–540. Springer, 2020.

[111] Y. Xian et al. Any-shot gin: Generalizing implicit networks for reconstructing novel classes. In *2022 International Conference on 3D Vision (3DV)*, pp. 526–535. IEEE, 2022.

[112] C.-Y. Wu et al. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9065–9075, 2023.

[113] Z. Huang et al. Zeroshape: Regression-based zero-shot shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10061–10071, 2024.

[114] J. J. Park et al. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

[115] Y. Ren et al. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16685–16695, 2023.

[116] Y. Liang et al. Retr: Modeling rendering via transformer for generalizable neural surface reconstruction. *Advances in neural information processing systems*, 36:62332–62351, 2023.

[117] L. Xu et al. C2f2neus: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18291–18301, 2023.

[118] Y. Na et al. Uforecon: generalizable sparse-view surface reconstruction from arbitrary and unfavorable sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5094–5104, 2024.

[119] Z. Wang et al. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pp. 57–74. Springer, 2024.

[120] F. Liu et al. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024.

[121] M. S. Sajjadi et al. Rust: Latent neural scene representations from unposed imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17297–17306, 2023.

[122] M. S. Sajjadi et al. Object scene representation transformer. *Advances in neural information processing systems*, 35:9512–9524, 2022.

[123] A. Safin et al. Repast: Relative pose attention scene representation transformer. *arXiv preprint arXiv:2304.00947*, 2023.

[124] M. Suhail et al. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pp. 156–174. Springer, 2022.

[125] Y. Du et al. Learning to render novel views from wide-baseline stereo pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4970–4980, 2023.

[126] N. Venkat et al. Geometry-biased transformers for novel view synthesis. *arXiv preprint arXiv:2301.04650*, 2023.

[127] T. Miyato et al. Gta: A geometry-aware attention mechanism for multi-view transformers. *arXiv preprint arXiv:2310.10375*, 2023.

[128] H. Jiang et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025.

[129] A. Van Den Oord et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[130] P. Esser et al. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

[131] R. Rombach et al. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14356–14366, 2021.

[132] J. Kulhánek et al. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pp. 198–216. Springer, 2022.

[133] K. Sargent et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9420–9429, 2024.

[134] T. Zhou et al. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[135] A. Liu et al. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14458–14467, 2021.

[136] R. Wu et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21551–21561, 2024.

[137] Y. Liu et al. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

[138] H. Weng et al. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023.

[139] J. Yang et al. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7079–7088, 2024.

[140] J. Ho et al. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[141] J. Xing et al. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

[142] N. Müller et al. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10258–10268, 2024.

[143] S. Wu et al. Genfusion: Closing the loop between reconstruction and generation via videos. In *CVPR*, 2025.

[144] J. Z. Wu et al. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *CVPR*, 2025.

[145] S. Zhang et al. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. 2025.

[146] P. Henzler et al. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9993, 2019.

[147] T. Nguyen-Phuoc et al. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7588–7597, 2019.

[148] T. H. Nguyen-Phuoc et al. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems*, 33:6767–6778, 2020.

[149] K. Schwarz et al. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[150] E. R. Chan et al. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5799–5809, 2021.

[151] M. Niemeyer and A. Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11453–11464, 2021.

[152] J. Gu et al. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[153] E. R. Chan et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

[154] S. Hyun and J.-P. Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *Advances in Neural Information Processing Systems*, 37:67987–68012, 2024.

[155] C. Smith et al. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow. *arXiv preprint arXiv:2306.00180*, 2023.

[156] S. Hong et al. Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20196–20206, 2024.

[157] X. Luo et al. Consistent video depth estimation. *ACM Trans. on Graphics*, 2020.

[158] W. Hu et al. Depthcrafter: Generating consistent long depth sequences for open-world videos. 2025.

[159] J. Shao et al. Learning temporally consistent video depth from video diffusion priors. *CVPR*, 2025.

[160] J. Kopf et al. Robust consistent video depth estimation. In *CVPR*, 2021.

[161] Z. Zhang et al. Structure and motion from casual videos. In *ECCV*, 2022.

[162] Z. Li et al. MegaSaM: accurate, fast, and robust structure and motion from casual dynamic videos. 2025.

[163] Z. Teed and J. Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *NIPS*, 2021.

[164] Y. Wang et al. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, 2024.

[165] L. Yang et al. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[166] J. Zhang et al. MonST3R: a simple approach for estimating geometry in the presence of motion. 2025.

[167] K. Xu et al. Das3r: Dynamics-aware gaussian splatting for static scene reconstruction. *arXiv.org*, 2024.

[168] R. Ranftl et al. Vision transformers for dense prediction. In *ICCV*, 2021.

[169] X. Chen et al. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv.org*, 2025.

[170] J. Lu et al. Align3r: Aligned monocular depth estimation for dynamic videos. 2025.

[171] J. Mai et al. Can video diffusion model reconstruct 4d geometry? *arXiv.org*, 2025.

[172] C. Cao et al. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv.org*, 2025.

[173] X. Ren et al. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025.

[174] Z. Jiang et al. Geo4d: Leveraging video generators for geometric 4d scene reconstruction, 2025.

[175] A. Team et al. Aether: Geometric-aware unified world modeling. *arXiv.org*, 2025.

[176] L. Jin et al. Stereo4d: Learning how things move in 3d from internet stereo videos. 2025.

[177] H. Feng et al. St4rtrack: Simultaneous 4d reconstruction and tracking in the world. *arXiv.org*, 2025.

[178] Z. Fan et al. Large spatial model: End-to-end unposed images to semantic 3d, 2024.

[179] X. Wang et al. Gsemsplat: Generalizable semantic 3d gaussian splatting from uncalibrated image pairs, 2024.

[180] A. Thai et al. Splattalk: 3d vqa with gaussian splatting, 2025.

[181] S. Zheng et al. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19680–19690, 2024.

[182] T. Kirschstein et al. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025.

[183] R. Khirodkar et al. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pp. 206–228. Springer, 2024.

[184] R. Murai et al. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16695–16705, 2025.

[185] D. Maggio et al. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

[186] G. Lu et al. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation, 2024.

[187] T. Yu et al. Manigaussian++: General robotic bimanual manipulation with hierarchical gaussian world model, 2025.

[188] A. X. Chang et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[189] M. Deitke et al. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

[190] Y. Cabon et al. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[191] L. Ling et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

[192] A. Dai et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

[193] X. Yu et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9150–9161, 2023.

[194] G. Baruch et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[195] P. Sun et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

[196] Y. Liao et al. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

[197] Y. Zheng et al. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19855–19865, 2023.

[198] W. Wang et al. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

[199] R. Jensen et al. Large scale multi-view stereopsis evaluation. In *CVPR*, pp. 406–413, 2014.

[200] X. Sun et al. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, pp. 2974–2983, 2018.

[201] L. Downs et al. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

[202] T. Wu et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 803–814, 2023.

[203] H. Xia et al. Rgbd objects in the wild: scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22378–22389, 2024.

[204] M. Deitke et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

[205] J. Shotton et al. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, pp. 2930–2937, 2013.

[206] J. Straub et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[207] J. Sturm et al. Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark. In *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, volume 13, pp. 6, 2012.

[208] A. Chang et al. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[209] M. Roberts et al. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

[210] N. Karaev et al. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13229–13239, 2023.

[211] C. Yeshwanth et al. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

[212] L. Mehl et al. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4991, 2023.

[213] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pp. 2041–2050, 2018.

[214] J. T. Barron et al. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.

[215] A. Knapitsch et al. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[216] T. Schops et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pp. 3260–3269, 2017.

[217] R. Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[218] M. Goesele et al. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 2402–2409. IEEE, 2006.

[219] Z. Zhang et al. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21508–21518, 2023.

[220] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.