From Few to Full: High-Resolution 3D Object Reconstruction from Sparse Views and Unknown Poses

G. Yao[1,2], S. Mavromatis[1] and J.-L. Mari[1]

[1]Aix Marseille Univ, CNRS, LIS, Marseille, France
[2]Retail-VR, Nantes, France

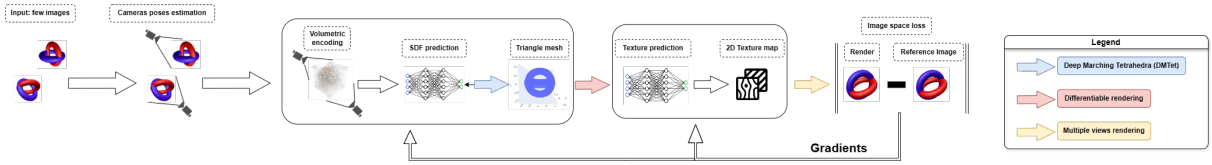

Figure 1: *Overview of the F2F Pipeline. We first predict camera poses and generate a tetrahedral feature volume from sparse input views. Secondly, we create a textured 3D mesh from the feature volume using DMTet, with all process optimized through differentiable rendering.*

Abstract

Recent progress in 3D reconstruction has been driven by generative models, moving from traditional multi-view dependence to single-image diffusion model based techniques. However, these innovative approaches often face challenges with sparse view scenarios, requiring known poses or template shapes, often failing in high-resolution reconstructions. Addressing these issues, we introduce the "F2F" (Few to Full) framework, designed for crafting high-resolution 3D models from few views and unknown camera poses, creating fully realistic 3D objects without external constraints. F2F employs a hybrid approach, optimizing both implicit and explicit representations through a unique pipeline involving a pretrained diffusion model for pose estimation, a deformable tetrahedra grid for feature volume construction, and an MLP (neural network) for surface optimization. Our method sets a new standard by ensuring surface geometry, topology, and semantic consistency through differentiable rendering, aiming for a comprehensive solution in 3D reconstruction from sparse views.

CCS Concepts

• ***Computing methodologies*** → *Sparse views; 3D reconstruction; Hybrid 3D representation; Differentiable rendering;*

1. Introduction

In real-world scenarios, it is common to have only a few views of an object available for 3D reconstruction. This problem is notably heightened by its ill-posed nature and the potential absence of complete object data in input images, and thus necessitates advanced techniques for accurately estimating unseen object features in alignment with available observations. This challenge has prompted the development of numerous methods aimed at reconstructing 3D objects from sparse views. Among these, surface-based methods have emerged as a foundational approach, leveraging the deformation of template shapes through differentiable rendering to align each input viewpoint with its corresponding rendering view. A notable implementation of this technique involves deforming predefined shapes, such as spheres or cuboids, using sparse views as a guide [ZYTR21]. However, this dependence on template shapes that correspond to input data considerably restricts the reconstruction of objects with complex or arbitrary topologies.

To overcome topology constraints, several studies have moved towards implicit-based methods, like optimizing SDF or neural radiance fields. For instance, given a sparse view, [MCCG23] predict a SDF representation by using volumetric rendering technique [MST*20]. [ZCC23] combines a NeRF-based technique with 2D diffusion model priors, optimizing the NeRF representation from sparse views while generating novel view images through a multiview-consistent diffusion model. Both methods still necessitate accurate camera poses and optimize only an implicit representation. [JJGZ24] predict camera poses directly in the 3D reconstruction pipeline. But all these approaches need a post-processing step using marching cubes algorithms to extract the final mesh, which introduces additional errors into the 3D reconstruction.

DOI: 10.2312/egp.20241045

Our method, "F2F" (Few to Full), is designed to directly optimize a surface representation without the constraints of topology and camera poses, outlined as follows:

- **Camera Pose Independence:** F2F eschews the need for predefined camera poses through the use of a pretrained diffusion model, enabling autonomous camera position estimation from sparse images.
- **Flexible Topology:** By adopting a hybrid 3D representation, F2F allows for seamless transitions from implicit functions to 3D meshes, enabling the generation of arbitrary topologies without relying on voxel grids.
- **Direct Surface Optimization:** Using differentiable marching tetrahedra, our method allows for the direct extraction and optimization of 3D surface meshes, eliminating the common requirement for postprocessing steps.
- **Semantic Shape Consistency:** F2F tackles the issue of sparse views by reconstructing unseen 3D parts through semantic shape consistency, comparing novel views from the reconstructed shape against semantic CLIP embeddings to ensure a thorough reconstruction from minimal data.

2. Method

The core methodology of F2F is depicted in Fig. 1, illustrating the process from input images to the final 3D reconstruction.

2.1. Pose Estimation and Feature Volume Construction

Using PoseDiffusion model, we estimate camera poses $\{C_i\}_{i=1}^{N}$ from sparse views $\{I_i\}_{i=1}^{N}$ and construct a feature volume on a deformable tetrahedral grid by projecting 2D features extracted with Dinov2 [ODM23] onto it, forming the basis for geometry reasoning.

2.2. Latent Vector and Prediction

For each vertex v in tetrahedral grid, we create a latent vector $z(v)$ as a concatenation of :

- A global semantic shape embedding derived from CLIP, summing embeddings for each input view image.
- A local semantic shape embedding obtained from the feature volume at vertex v.
- A Positional encoding used to capture the spatial features of each vertex v.

We then predict the signed distance function (SDF) value $s(v)$ and a deformation vector $\delta(v)$ using an MLP:

$$\{s(v), \delta(v)\} = MLP(z(v))$$

2.3. Surface Extraction and Optimization

The predicted values enable us to extract a triangular mesh from the tetrahedral grid using the marching tetrahedra technique while predicting a 2D texture map using a second MLP. We define several loss functions to directly optimize the surface mesh and texture map.

- **Photometric Loss**: We render RGB images and masks $\{\hat{I}, \hat{M}_i\}_{i=0}^{N-1}$ from the mesh, computing photometric L2 loss $L_{rgb} = \|I - \hat{I}\|_2$ and mask L2 loss $L_{mask} = \|M - \hat{M}\|_2$.
- **Semantic Shape Consistency Loss**: We render a set of sparse novel viewpoints $\hat{I}_{novel}$ from the reconstructed object, obtain CLIP embeddings for these images, and compute the semantic loss by comparing these new embeddings to the global CLIP embedding: $L_{ssc} = \left\| \sum embed(\hat{I}_{novel}) - \sum embed(I) \right\|_2$
- **Regularization Loss**: We also define a regularization loss for the predicted vertex deformation to avoid artifacts.

The total loss is a weighted combination of these losses, formulated as: $L_{total} = \lambda_{rgb} L_{rgb} + \lambda_{mask} L_{mask} + \lambda_{ssc} L_{ssc} + \lambda_{reg} L_{reg}$

3. Results

We show in Fig. 2 novel views from 3D models reconstructed by F2F. We can mainly observe that F2F generate triangular meshes with arbitrary topologies.

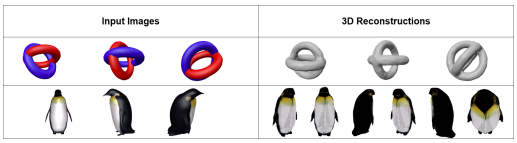

Figure 2: *Examples of 3D reconstruction results from a set of few 2D synthetic images with known poses.*

4. Conclusion

Our goal is to propose a method for high-resolution 3D reconstruction from sparse views, addressing the complexities of unknown camera poses and arbitrary topologies. Through a hybrid representation and surface-focused optimization, we demonstrate that our pipeline effectively produces 3D models from few 2D synthetic data with known poses. In future work, we will extend testing to broader datasets, incorporating real-world data without predefined poses, and explore how semantic shape consistency can further enhance neural network-based 3D reconstruction.

References

[JJGZ24] JIANG H., JIANG Z., GRAUMAN K., ZHU Y.: Few-view object reconstruction with unknown categories and camera poses. *International Conference on 3D Vision (3DV)* (2024). 1

[MCCG23] MU T.-J., CHEN H., CAI J., GUO N.: Neural 3d reconstruction from sparse views using geometric priors. *Computational Visual Media 9* (2023), 687 – 697. 1

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM 65* (2020), 99–106. 1

[ODM23] OQUAB M., DARCET T., MOUTAKANNI T.: Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193* (2023). 2

[ZCC23] ZOU Z.-X., CHENG W., CAO Y.-P.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. *AAAI* (2023). 1

[ZYTR21] ZHANG J. Y., YANG G., TULSIANI S., RAMANAN D.: Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Neural Information Processing Systems* (2021). 1