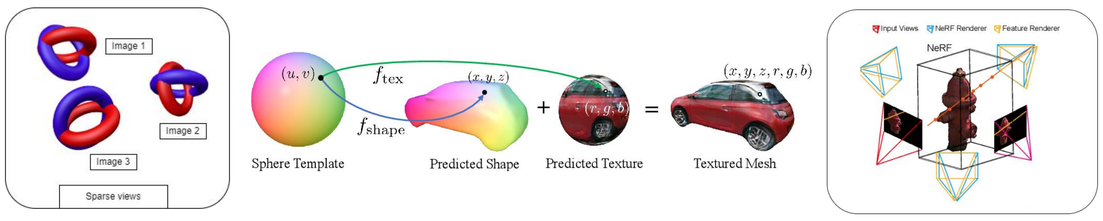
EUROGRAPHICS CONFERENCE 2024

FROM FEW TO FULL: HIGH-RESOLUTION 3D OBJECT RECONSTRUCTION FROM SPARSE VIEWS AND UNKNOWN POSES

G. Yao , S. Mavromatis and J.-L. Mari
Aix Marseille Univ, CNRS, LIS, Marseille, France
Retail-VR, Nantes, France

PROBLEM

In real-world scenarios, it is common to have only a few views of an object available for 3D reconstruction. This issue is particularly daunting due to its ill-posed nature, compounded by the fact that certain portions of the 3D model may not be captured in the available input images. Consequently, there is a critical need for techniques capable of accurately inferring or "hallucinating" the unseen aspects of an object's shape and texture in a manner that remains consistent with the observed views.



RELATED WORK

Several approaches have been proposed:

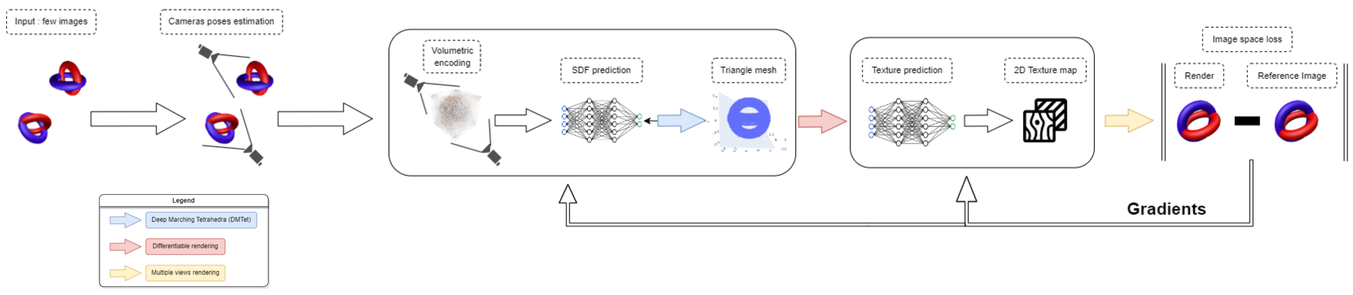
- **Surface based methods**: leveraging the deformation of template shapes (spheres, cuboids, etc.) through differentiable rendering to align each input viewpoint with its corresponding rendering views [7].
- **Implicit-based methods:** optimize a signed distance function (SDF) as in [2] and [1], or a neural radiance field [6] using volumetric rendering [3].

Limitations: Surface-based methods require fixed topology, hindering arbitrary 3D model reconstruction while Implicit-based methods optimize only implicit forms, necessitating postprocessing (e.g. marching cubes) for mesh extraction, introducing potential errors and low resolution. Both approaches require accurate camera poses,

OVERVIEW

We introduce our method 'F2F' (Few to Full) to directly optimize a surface representation, giving sparse object views without the constraints of topology and camera poses. 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.

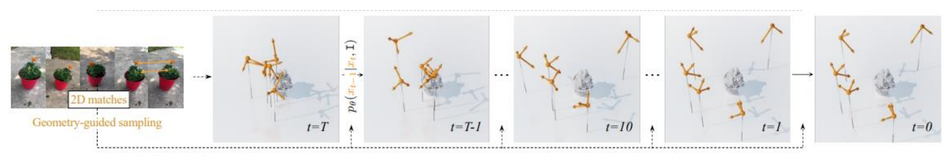
METHODOLOGY



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

- **Poses estimation and Feature volume construction**

We use PoseDiffusion model [5] to estimate camera poses from sparse views, then we construct a feature volume on a deformable tetrahedral grid by projecting 2D features extracted with Dinov2 [4] onto it, forming the basis for geometry reasoning.



- **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 embeddings
- a local semantic shape embedding from the feature volume at vertex v
- a positional encoding used to capture the spatial features of each vertex v

We predict signed distance function (SDF) value s(v) and deformation vector δ(v) using MLP:

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

- **Surface Extraction and Optimization**

With {s(v),δ(v)} predicted, we extract triangular mesh using marching tetrahedra. Then we predict a 2D texture map by using a second MLP. To directly optimize the surface mesh, we render multi-view images of predicted mesh, and define several losses:

Photometric loss **:** computing L2 loss to compare predicted images to ground truth images as $L_{rgb} = \left\|I - \hat{I}\right\|_2$ and $L_{mask} = \left\|M - \widehat{M}\right\|_2$

Semantic Shape Consistency loss **:** computing L2 loss to compare predicted images CLIP embeddings to the global CLIP embeddings as $L_{\text{ssc}} = \left\|\sum embed(\hat{I}_{novel}) - \sum embed(I)\right\|_2$

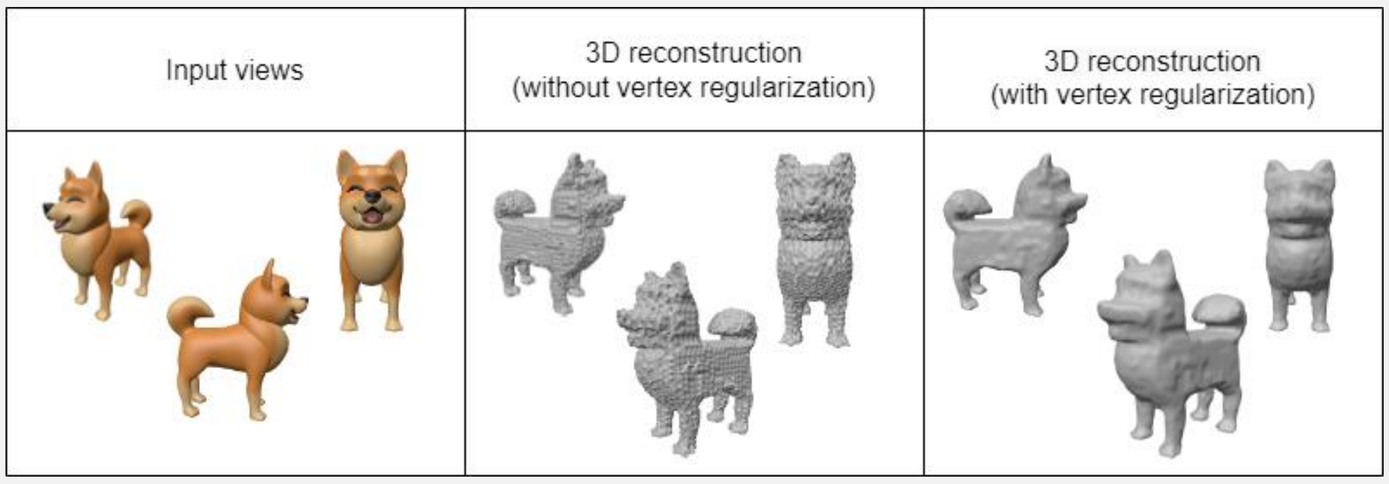
Regularization loss **:** Applied to vertex deformation to avoid artefacts.

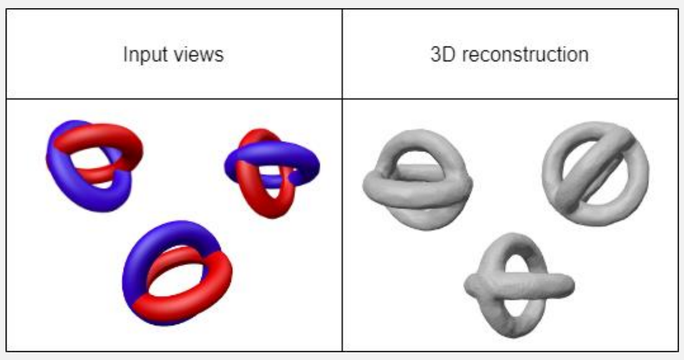
Total loss is weighted combination of these losses: $L_{total} = \lambda_{rgb} l_{rgb} + \lambda_{mask} L_{mask} + \lambda_{ssc} L_{ssc} + \lambda_{yeg} L_{reg}$

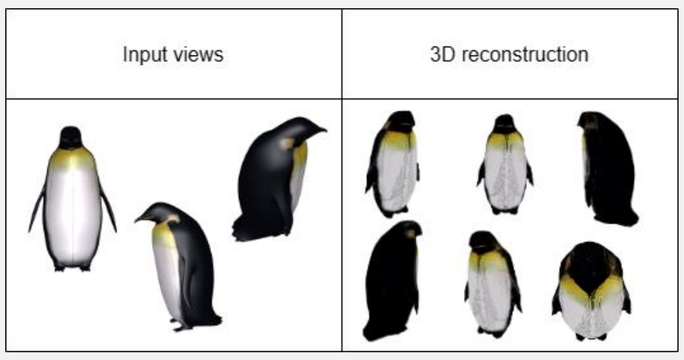
RESULTS

We test our method on synthetic dataset. For each object, we render from 6 to 10 synthetic 2D views with predefined camera poses. Then we use these set of sparse views to reconstruct 3D models. We mainly observe that F2F generates triangular meshes with arbitrary topologies.

In conclusion, we aim 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 a few 2D synthetic data with known poses. In future work, we will extend testing to more extensive datasets, incorporating real-world data without predefined poses, and explore how semantic shape consistency can further enhance neural network-based 3D reconstruction.







AFFILIATIONS

Aix Marseille université
retail VR
LIS Laboratoire d'Informatique & Systèmes
G-Mod Geometric Modeling

REFERENCES

[1] 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).
[2] 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.
[3] 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.
[4] OQUAB M., DARCET T., MOUTAKANNI T.: Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193* (2023).
[5] WANG J., RUPPRECHT C., NOVOTNY D.: PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV* (2023)
[6] ZOU Z.-X., CHENG W., CAO Y.-P.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. *AAAI* (2023).
[7] 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).