3D reconstruction with single/multi-view images
Detail Project Plan

Jiang, Zeyu
jzy0415@connect.hku.hk

Zhao, Hengshuang (supervisor)
hszhao@cs.hku.hk

1. Background

Humans could infer the geometry and texture of a 3D object merely from a few images, relying on their strong prior knowledge built over a lifetime of visual experience. However, enabling machines to perform 3D reconstruction from limited input views remains an open challenge with many applications across robotics, autonomous vehicles, augmented reality and more.

1.1 Pure 3D reconstruction

Classic methods based on multi-view stereo (MVS) [1, 2] can reconstruct 3D models from multiple images taken from known viewpoints. However, they struggle with textureless surfaces, lighting variations, and thin structures. MVS also requires many input views captured in a controlled setting.

Recently, an increasing number of methods have adopted implicit neural representations for modeling 3D scenes, demonstrating strong reconstruction performance and high fidelity results [3, 4, 5, 6]. However, they also require large number of input images with known poses and may produce blurry output when facing areas unseen.

1.2 Diffusion Models

As diffusion models [7, 8] showed remarkable performance in 2D images generation, many works seek to transfer this ability for 3D tasks.

DreamFusion [9], as a pioneer work, utilizes a pretrained text-to-image diffusion model to generate 3D models from text prompts. Given the inspiration, some works [10, 11, 12, 13] adapt the pipeline and focus on optimizing a text-inversion model to synthesize novel views from single image for reconstruction.

Another pioneer work Zero-1-to-3 [14], introduces an architecture of view-conditioned diffusion model, which could generate novel views from any angle given the single image input. Followed by this, more recent works [15, 16] model the multi-view images generated by diffusion model for reconstruction as a joint distribution to maintain 3D consistency. Moreover, this line of works focus more on creation of object matches with input image rather than reconstruction.

There are also works like SparseFusion [17] optimizing the diffusion model with reconstruction process to keep the consistency of geometry. However, this kind of model normally won't work in the open world and may require categorical prior.

In summary, major limitations of existing methods include reliance on known poses, controlled capture, category-specific training, and degradation with sparse inputs. The ability to reconstruct high-quality 3D geometry from one or a handful of in-the-wild images remains an open problem.

2. Objective

The goal is to develop a scalable framework for 3D reconstruction that can leverage either single or multiple input images. Key objectives include:

- Develop a model that can reconstruct 3D geometry from single image.
- Enable multi-view 3D reconstruction where additional images can incrementally improve the reconstruction.
- Support input images without categorical prior, masks or poses.
- Evaluate reconstruction quality compared to state-of-the-art methods in closed-world benchmarks quantitively.
- Qualitatively assess plausibility of completions with other state-of-the-art methods.
- Analyze tradeoffs between single vs. multi-view reconstruction in terms of accuracy and runtime.

The aim is a flexible framework that utilizes single or multiple images to produce complete 3D models. Additional views should enhance reconstruction quality when available.

3. Methodology

This project will develop a framework for few-shot single and multi-view 3D reconstruction of arbitrary objects. The approach will build upon recent advances in view-conditioned diffusion models and neural renderings. Following is the rough discussion on methodology.

3.1 Input preprocessing

As a first step, any available input images will be preprocessed to generate additional pseudo-views. This can provide more supervision for the model and emulate a multi-view capture setup from a single photo. Data augmentations like random cropping and flipping will be applied to extract plausible alternative perspectives. With camera poses and masks unknown, approximate viewpoints and masks will also be estimated using off-the-shelf pretrained models.

3.2 View-conditioned diffusion model

We may still base on the pretrained Zero-1-to-3 [14] model for novel view synthesize. To keep the 3D consistency, we may learn from the joint distribution approach [15, 16] and distillation approach [9, 17]. We may also redesign the pipeline to make the model could condition on incremental input images for enhancement.

3.3 Reconstruction

Given the recognition of NeRF [3] and its following improvements [4, 5, 6], we plan to adopt the latest method while adding features such that the reconstruction process may guide the generation of novel views.

4. Schedule and Milestones

The following table summarize the schedule and milestones of this final year project.

4.1. Preparation and literature review (Sep 1, 2023 - Oct 15, 2023)

In the first and a half month, we'll focusing on mostly literature review. Checking with latest work in relevant areas and try to reproduce their results. Need to select and determine the method for each step mentioned in the methodology part above. Also, should check with the environment and framework (PyTorch) setup. Milestones of this stage listed as follows.

- Summary of literature review
- Complete setup environment
- A more concrete version of methodology

4.2. Prelinminary implementation of the framework (Oct 16, 2023 - Nov 30, 2023)

Implement a prelinmary version of the framework, while also working on the interim report. The prelinmary implementation may assume the input images are with masks and camera positions. The key point is to implement and finetune the diffusion and reconstruction network. Milestones of this stage listed as follows.

- Prelinminary implementation of the framework
- Test result of prelinmary framework
- Draft of interim report

4.3. Improvement of implementation (Dec 1, 2023 - Jan 15, 2023)

In this stage, try to improve the performance of original implementation with consideration on how to handle the wild images input. Check with latest work to see if there's any novel method. Also, complete the interim report. Milestones of this stage listed as follows.

- Complete implementation of the framework
- Finalized version of interim report
- Interim presentation

4.4. Finalization and future work (Jan 16, 2023 - Apr 15, 2023)

During this period, we should finalize all the implementation and report. Future work include deployment an online demo for the framework, optimize the network to be mobile-capable, etc.

- Final implementation
- Final report
- Materials for exhibition

References

[1] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 1

[2] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2012. 1

[3] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng.

Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[4] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 1, 2

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1, 2

[6] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, 2023. 1, 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[9] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 2

[10] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 1

[11] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 1

[12] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior, 2023. 1

[13] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1

[14] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 1, 2

[15] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 2

[16] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. *arXiv preprint arXiv:2303.17905*, 2023. 1, 2

[17] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 1, 2