Chair of Computer Aided Medical Procedures
TUM School of Computation, Information and Technology
Technical University of Munich

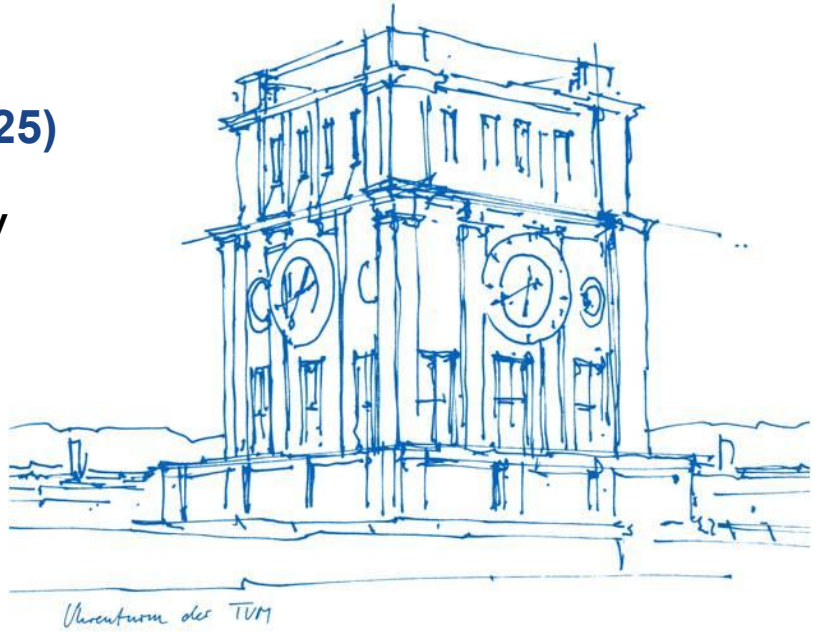
Visual Grounded Open-Vocabulary Object Pose Estimation

Advanced Topic in 3D Computer Vision (SS25)

Haoliang Huang, Yung Jhang Hou, Namozjon Ostonaev

School of Computation, Information and Technology
Technical University of Munich

Munich, 24. July 2025



Outline

- Pipeline evolution: initial plan vs. real-world constraints
- RGB→RGB-D conversion
- Generation models: replacing the mesh input
- Evaluation: comparing generated meshes
- Future work
- References

Motivation

Motivation

Can we estimate **depth**, reconstruct a **3D mesh**, and recover **6D object pose** — all from a single RGB image?

6D Object Pose

- The **camera is the origin** of the coordinate system (0, 0, 0), facing along the Z-axis.
- The **6D pose** of an object tells:
  a. **Where the object is** (translation: X, Y, Z)
  b. **How it's rotated** (rotation: α, β, γ or roll, pitch, yaw)

6D Pose = 3D Position (X, Y, Z) + 3D Rotation (α, β, γ)

Initial Pipeline (Any6D)

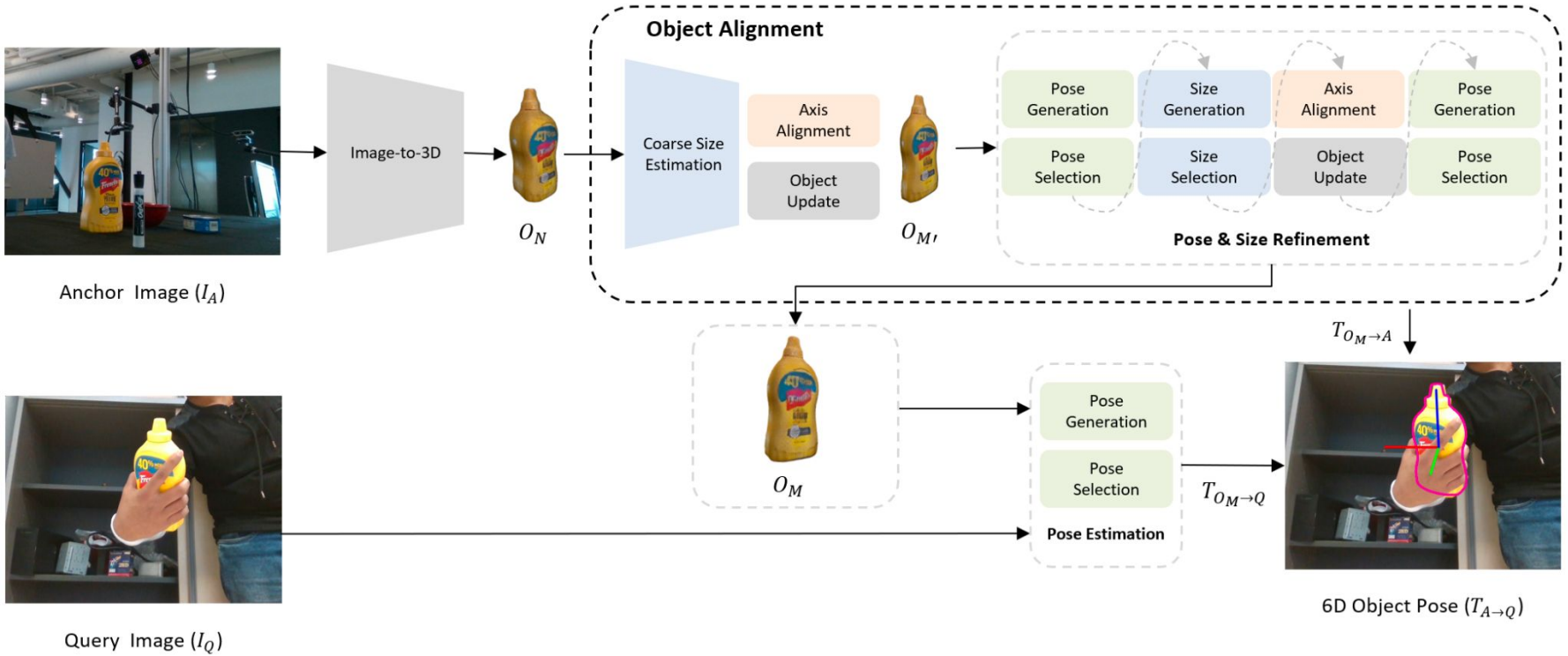
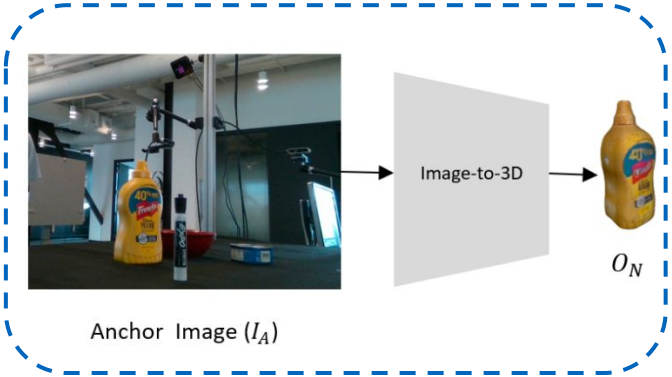


Figure. Overview of the Any6D framework for model-free object pose estimation.

Lee, Taeyeop and Wen, Bowen and Kang, Minjun and Kang, Gyuree and Kweon, In So and Yoon, Kuk-Jin. {Any6D}: Model-free 6D Pose Estimation of Novel Objects, In Proceedings of the IEEE/CVF international conference on computer vision,2025

Mesh Generation of Any6D



InstantMesh

Zero123++

Figure. Explanation of what is behind the generation model of Any6D.

Lee, Taeyeop and Wen, Bowen and Kang, Minjun and Kang, Gyuree and Kweon, In So and Yoon, Kuk-Jin. {Any6D}: Model-free 6D Pose Estimation of Novel Objects, In Proceedings of the IEEE/CVF international conference on computer vision,2025.

Yu, Xiaoxu and Liu, Jin and He, Bo and Xu, Yan and Chen, Xiaolong and Zeng, Bing and Li, Xin. {InstantMesh}: Efficient 3D Gaussian Splatting via Instant Content-Aware Meshing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

Li, Ruohan and Chen, Xiang and Han, Jiaming and Li, Jiawei and Tang, Fang and Zhang, Fan and Yin, Ruoyu and Liu, Khawla and Xia, Hao. {Zero123++}: A Single-Image to 3D Model Generator. arXiv preprint arXiv:2311.16450, 2023.

Problems!

- Mesh generation was not good enough
- RGB-**D** image was required

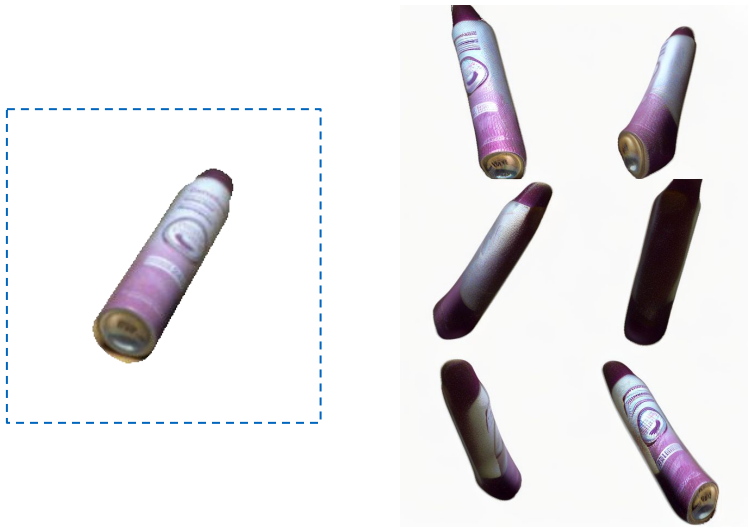


Figure. input target and distorted multiview generation

Updated Pipeline

Replace the generation model

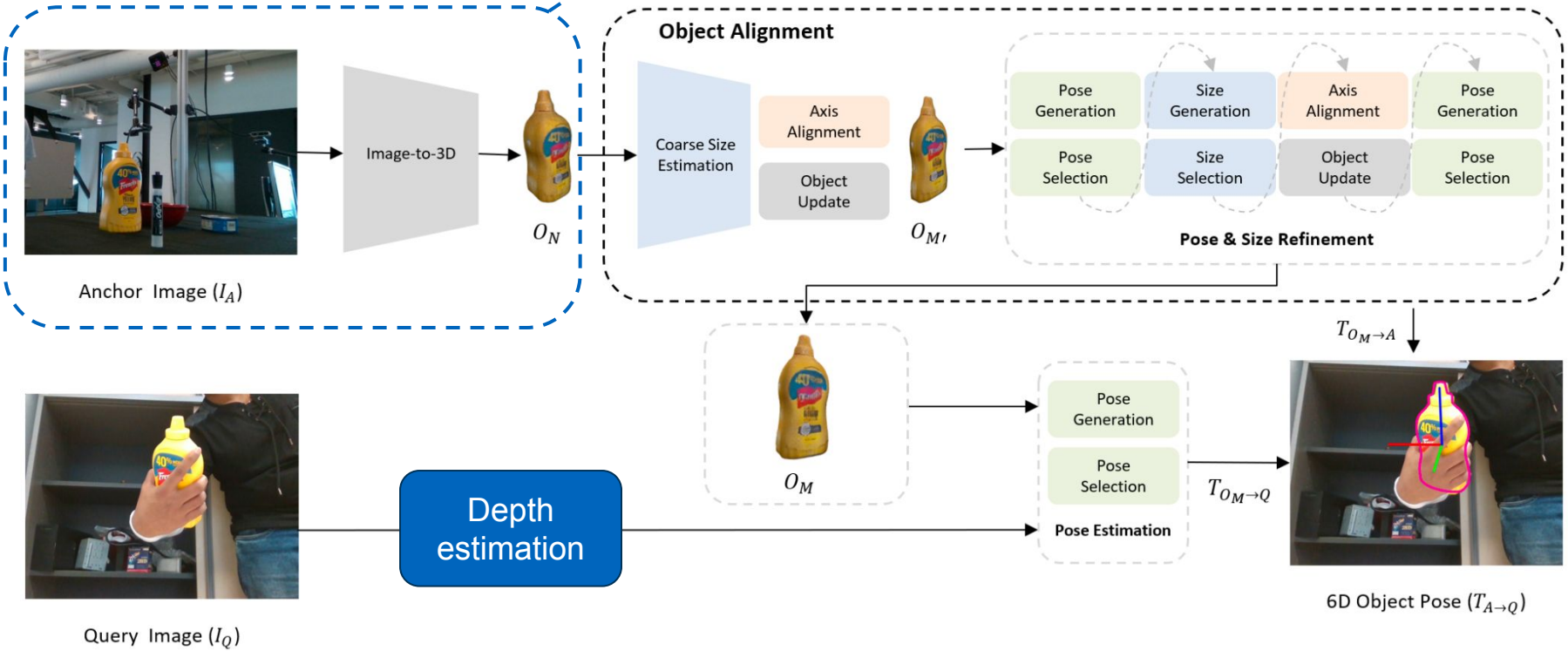
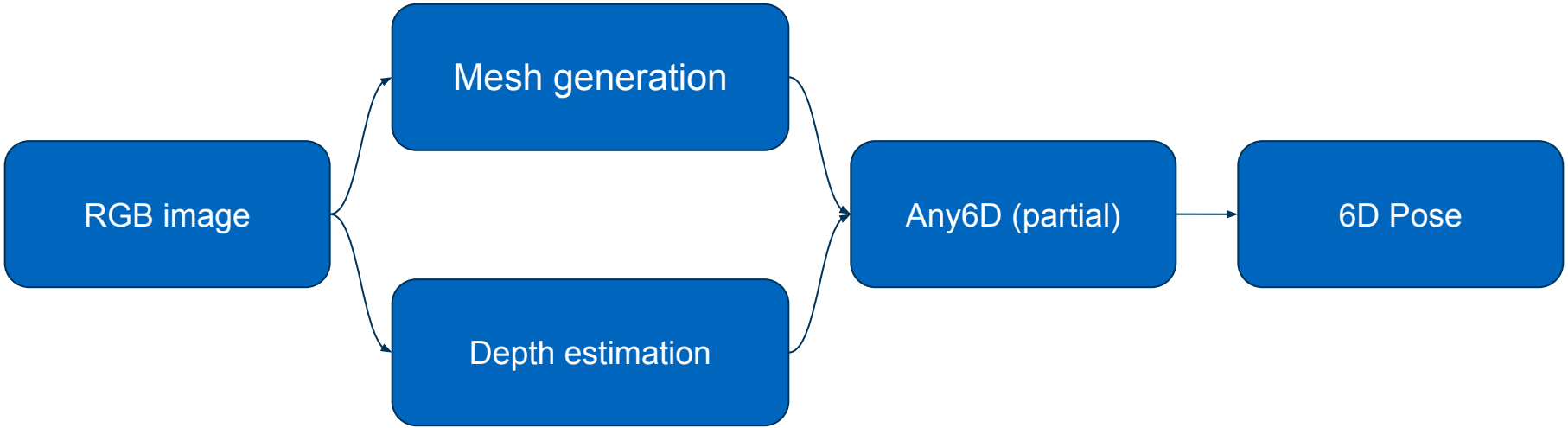


Figure. Overview of the Any6D framework for model-free object pose estimation.

Lee, Taeyeop and Wen, Bowen and Kang, Minjun and Kang, Gyuree and Kweon, In So and Yoon, Kuk-Jin. {Any6D}: Model-free 6D Pose Estimation of Novel Objects, In Proceedings of the IEEE/CVF international conference on computer vision,2025

Updated Pipeline (simplified)



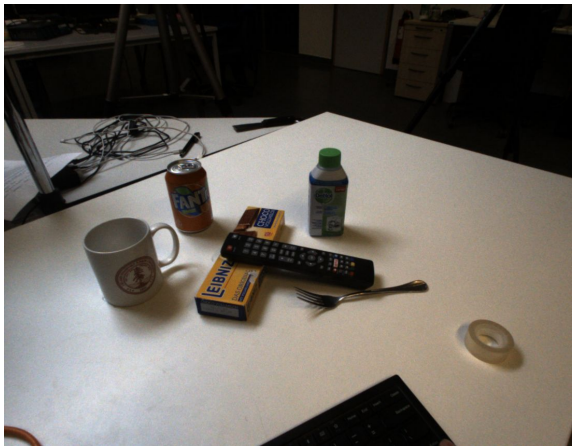
Experiments

Experiments

- Depth estimation
- Preprocessing
- Multiview Generation Models
- Single Image to 3D Generation Models
- Evaluation Metrics

Dataset: HouseCat6D

- Indoor images, including 10 scenes, shoots from several angles for each scene
- Segmentation masks
- Depth maps
- Camera poses
- Object poses



Depth Estimation

VGGT: Visual Geometry Grounded Transformer

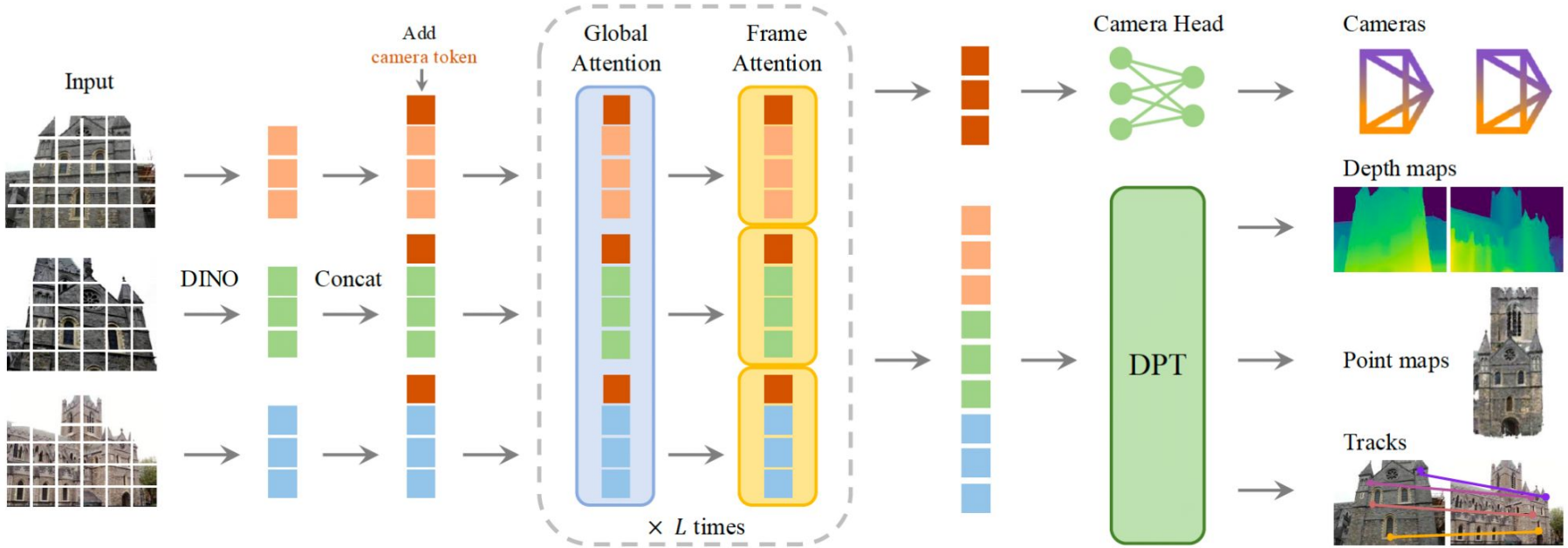


Figure. VGGT Architecture Overview.

Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. *arXiv.org*, arxiv.org/abs/2503.11651

VGGT: Results from the paper

*Figure 3. Reconstruction of the Egyptian pyramid with **VGGT***

Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. *arXiv.org*, arxiv.org/abs/2503.11651

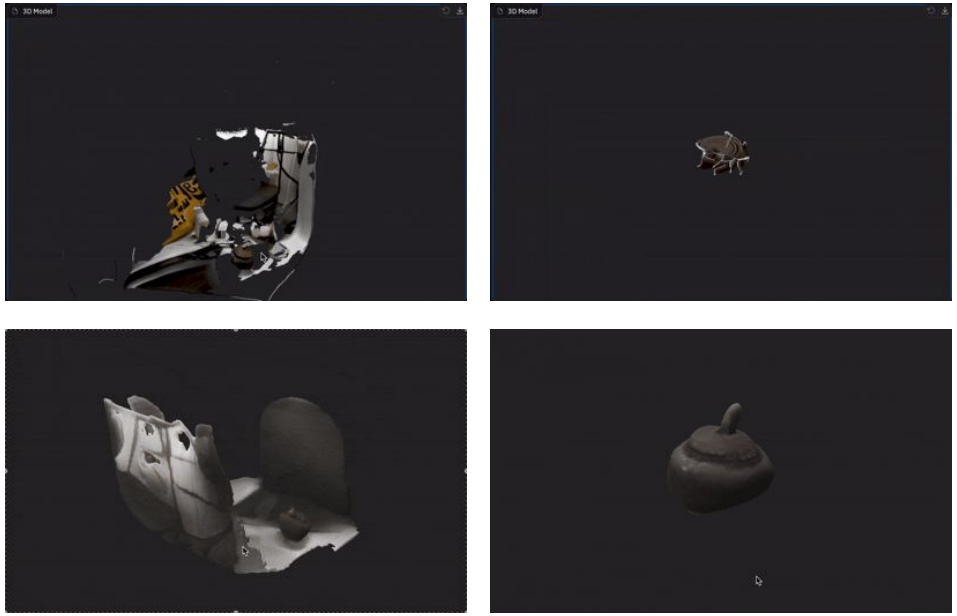
VGGT: Results from housecat6D

Target Image

*Figure. Given a target image from **housecat6D** dataset, 6 images are generated by **Zero123++**. Here are the results of **VGGT** point cloud, and **InstantMesh** mesh, both reconstructed from these 6 images.*

VGGT

InstantMesh



Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. *arXiv.org*, arxiv.org/abs/2503.11651

Yu, Xiaoxu and Liu, Jin and He, Bo and Xu, Yan and Chen, Xiaolong and Zeng, Bing and Li, Xin. {InstantMesh}: Efficient 3D Gaussian Splatting via Instant Content-Aware Meshing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

Li, Ruohan and Chen, Xiang and Han, Jiaming and Li, Jiawei and Tang, Fang and Zhang, Fan and Yin, Ruoyu and Liu, Khawla and Xia, Hao. {Zero123++}: A Single-Image to 3D Model Generator. arXiv preprint arXiv:2311.16450, 2023.

VGGT: Visual Geometry Grounded Transformer

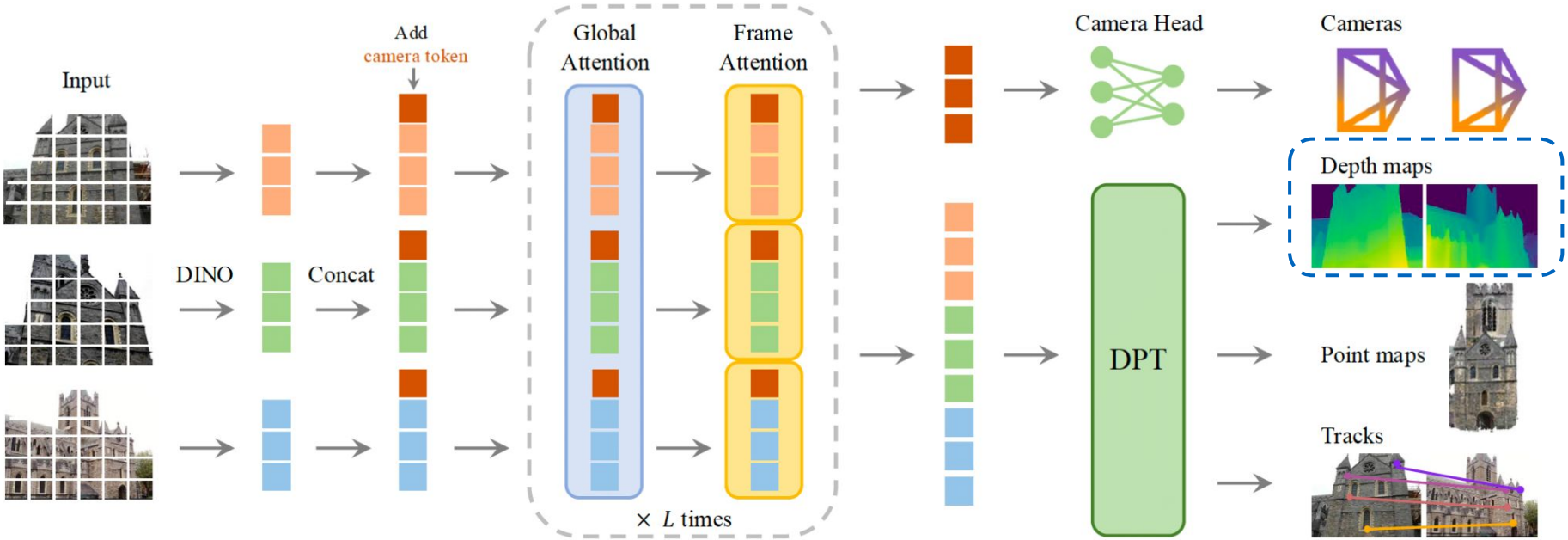


Figure. VGGT Architecture Overview.

Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. *arXiv.org*, arxiv.org/abs/2503.11651

VGGT: Depth Estimation

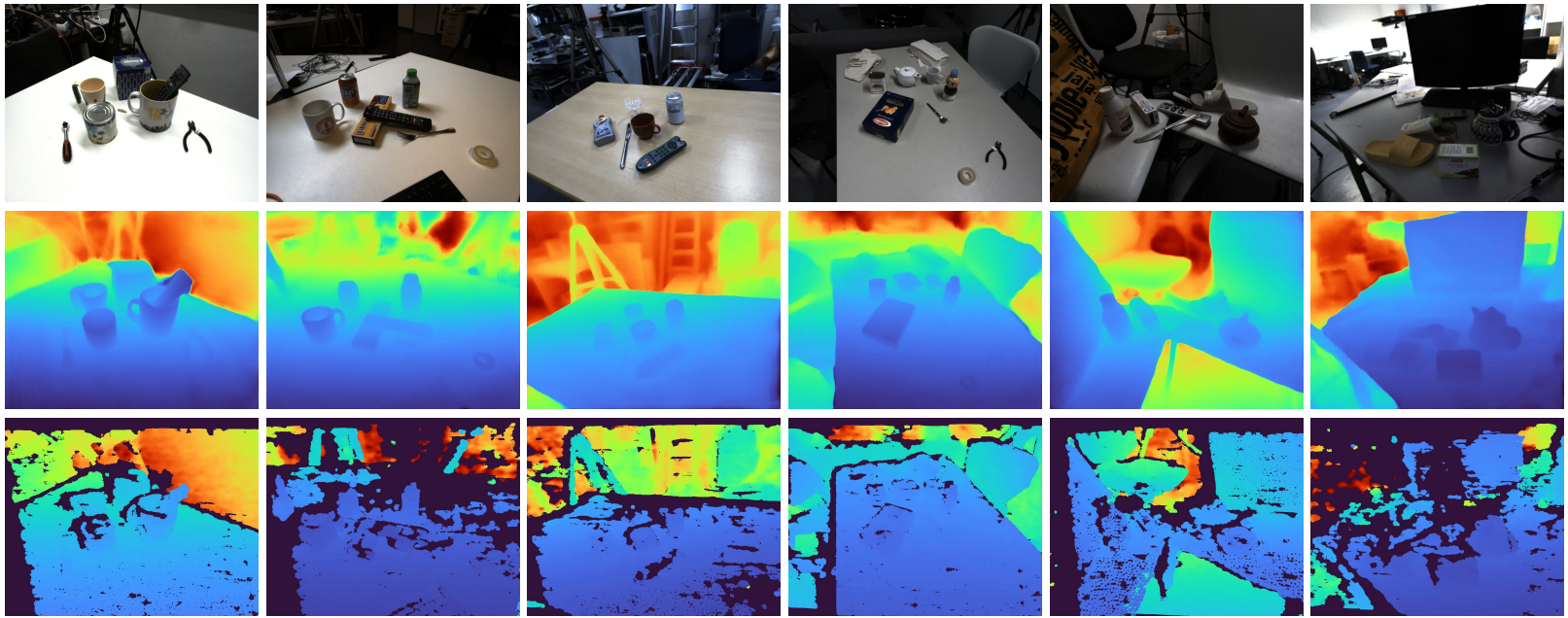


*Figure. Given **housecat6D** dataset, the first row are the target images, second row are the estimated depths by **VGGT**, and the third row are the ground truth depths.*

Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. *arXiv.org*, arxiv.org/abs/2503.11651

3D Mesh Generation

Segmentation Methods

Grounded-Sam-2: Grounding DINO + SAM2 (Pipeline)

- We need **unique description** for this target in a scene.
- We **cannot use overly specific names**. It would **confuse Grounded-Sam-2's segmentation**.

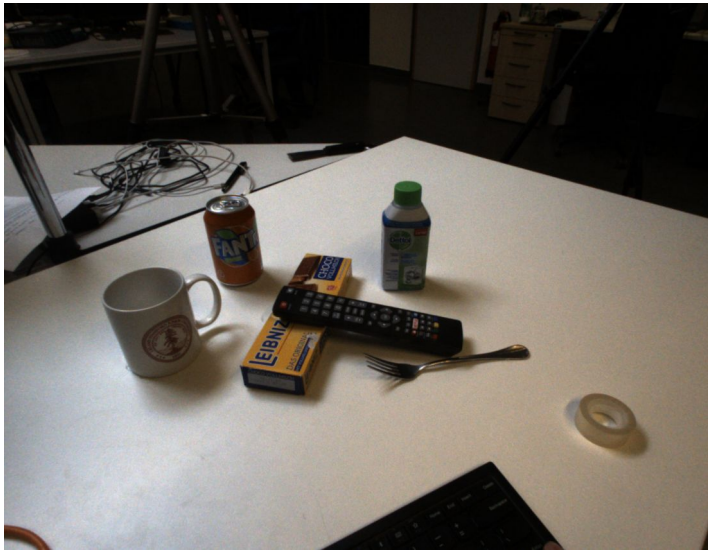


Figure: Original Scene Input

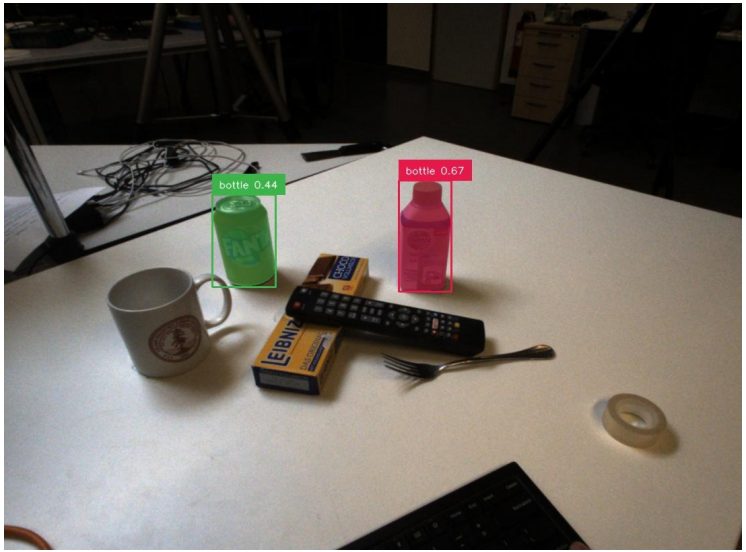


Figure: Selection Result from Grounded-Sam-2

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., ... & Zhang, L. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159.
Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., ... & Zhang, L. (2024). Grounding dino 1.5: Advance the" edge" of open-set object detection. arXiv preprint arXiv:2405.10300.
Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., ... & Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

Preprocess methods

Tried different preprocess method: **Center, Padding(0%-100%), Rembg**

Figure: Imput images form padding 0% - 100%

Figure: Results form padding 0% - 100%

Preprocess methods

Centered + 30% padding has best average performance, while rembg provides more stable results

Chamfer Distance Comparison between different Preprocess

Method	Grounded-Sam2	Grounded-Sam2 + Rembg
bbox	0.654	0.572
10% padding	0.608	0.567
20% padding	1.190	0.578
30% padding	**0.538**	0.569
40% padding	1.444	0.604
50% padding	0.583	0.567
60% padding	1.232	0.568
70% padding	1.964	**0.564**
80% padding	0.603	0.611
90% padding	0.617	0.569
100% padding	0.571	0.605

Figure: Center + 30% padding

Figure: Rembg-Tool

U^2-Net generation

Citation: https://github.com/danielgatis/rembg
Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. Pattern recognition, 106, 107404.

So we choose: **Center + 30% padding + Rembg-Tool,**

- padding size does not matter so much

Generation models: Multiview Generation

Baseline: InstantMesh

fine tuned Zero123++
(on white background)

Reconstruction Model

Single RGB Image → **Multiview RGB Images** → **3D Model**

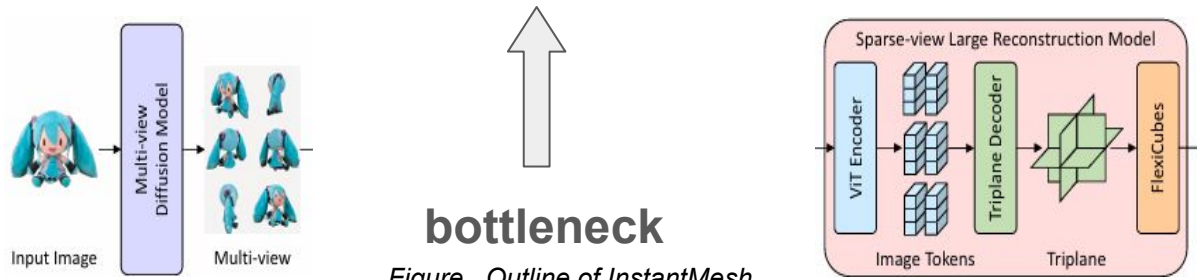


Figure . Outline of InstantMesh

Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., & Shan, Y. (2024). Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191.

Generation models: Multiview Generation

EscherNet

Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., & Davison, A. J. (2024). Eschernet: A generative model for scalable view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9503-9513).

SV3D (Video Generation)

Voleti, V., Yao, C. H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., ... & Jampani, V. (2024, September). Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In European Conference on Computer Vision (pp. 439-457). Cham: Springer Nature Switzerland.

MVGenMaster (Video Generation)

Cao, C., Yu, C., Liu, S., Wang, F., Xue, X., & Fu, Y. (2025). MVGenMaster: Scaling Multi-View Generation from Any Image via 3D Priors Enhanced Diffusion Model. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 6045-6056).

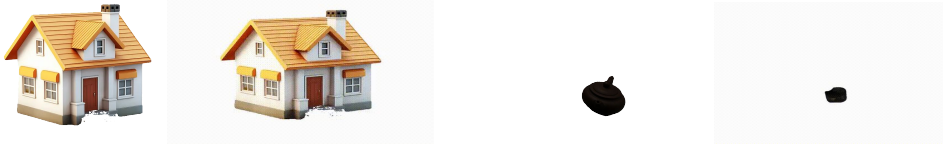


Figure:
Line1: Eschernet result (image, result)
Line2: SV3D result (image, result)
Line2: MVGenMaster result (image, result)

Generation models: Multiview Generation

Figure:
Line1: input image
Line2: Era3D result
Line3: InstantMesh result

Era3D

InstantMesh
Best Case

Li, P., Liu, Y., Long, X., Zhang, F., Lin, C., Li, M., ... & Guo, Y. (2024). Era3d: High-resolution multiview diffusion using efficient row-wise attention. Advances in Neural Information Processing Systems, 37, 55975-56000.
Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., & Shan, Y. (2024). Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191.

Generation models: Multiview Generation

Conclusion:

- Results are not better than InstantMesh.

Comment:

- Multiview generation models are outdated.
- Before 2025, multiview generation is necessary for 3d mesh generation
- But in 2025, multiview is no longer necessary as the separate part of image-to-3d model.
- Current Model are tend to adopt an end-to-end design.

Generation models

MIDI-3D

Image -> 3D Instances

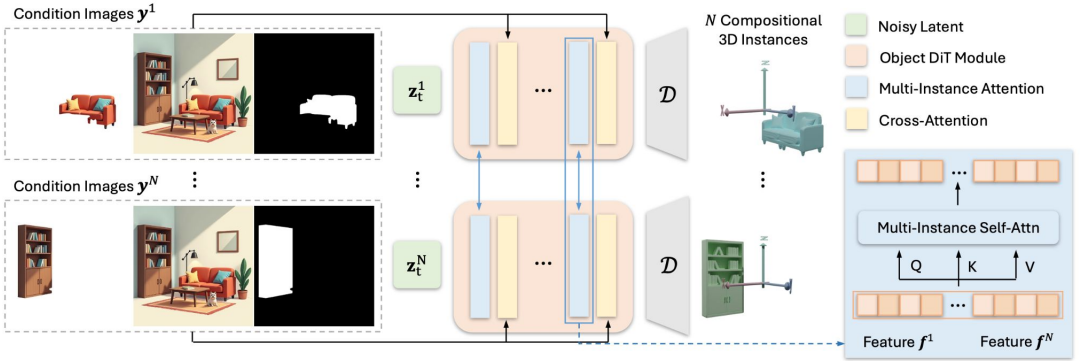


Figure . Outline of Midi-3D

Huang, Z., Guo, Y. C., An, X., Yang, Y., Li, Y., Zou, Z. X., ... & Sheng, L. (2025). Midi: Multi-instance diffusion for single image to 3d scene generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 23646-23657).

Hi3DGen

Image -> normal map -> geometry

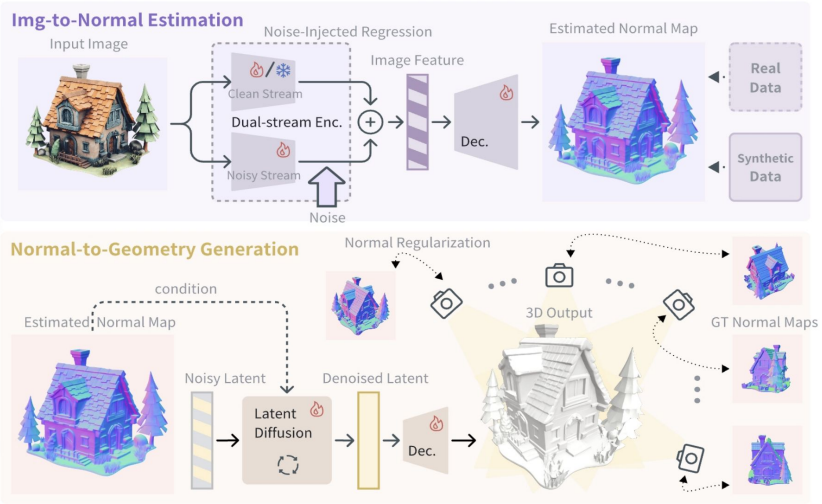


Figure . Outline of Hi3DGen

Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., ... & Han, X. (2025). Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236, 3, 2.

Generation models

Hunyuan3D 2.5

Image -> 3D mesh

Image + 3D mesh -> Texture

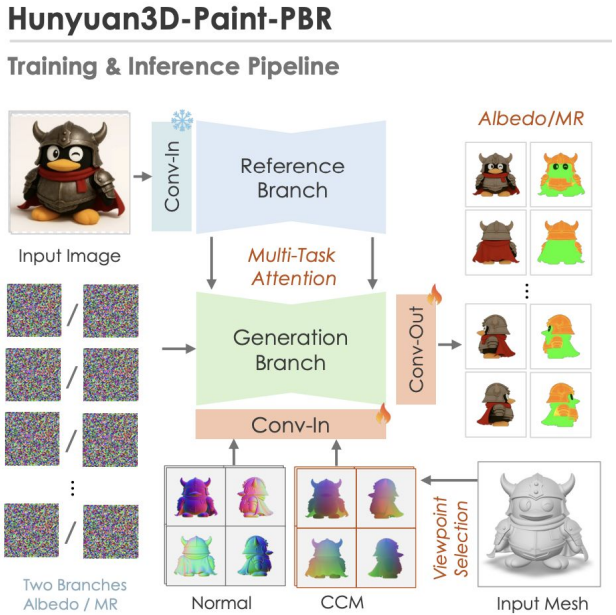


Figure . Outline of Hunyuan3D 2.5

TRELLIS

Train: 3D model -> latent space -> 3D model

Inference: latent value + image -> 3D model

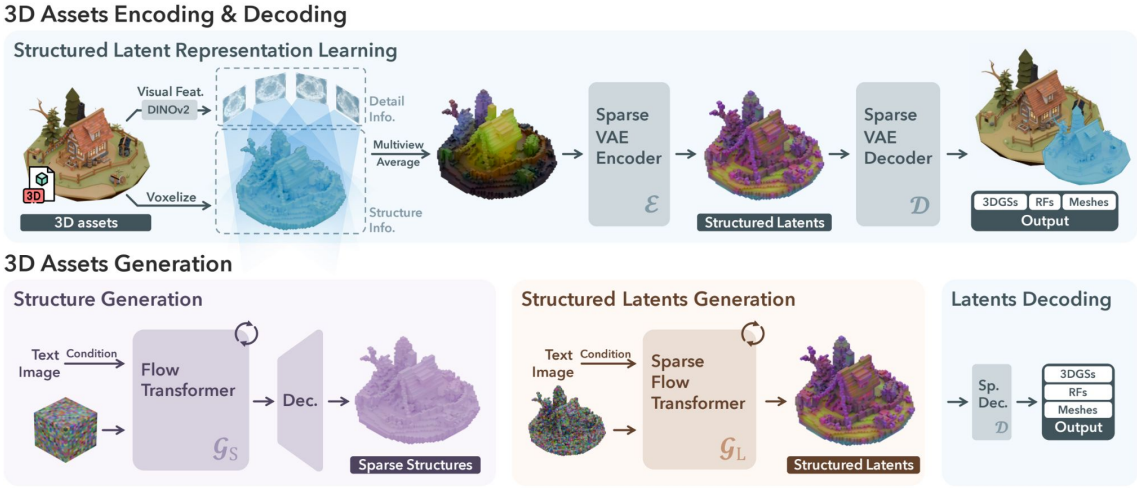


Figure . Outline of TRELLIS

Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., ... & Guo, C. (2025). Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details. arXiv preprint arXiv:2506.16504.

Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., ... & Yang, J. (2025). Structured 3d latents for scalable and versatile 3d generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 21469-21480).

Generation models: Image-to-3d Model

Tried different entire Generation Model:

- **New released model in 2025**
  - CraftsMan, Hunyuan3D 2.5, Hi3dGen, TRELLIS, MIDI-3d.

Comments

- **InstantMesh** and **CraftsMan** is traditional pipeline: **image -> multiview -> 3d mesh**
- **Hunyuan** specifically designed for a best **high-quality material generation**.
- **Hi3dGen** specifically designed for a best **high-quality geometry generation**.
- **TRELLES** and **MIDI-3D** are used to **generate the whole scene**.

Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., & Shan, Y. (2024). Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191.
Li, W., Liu, J., Yan, H., Chen, R., Liang, Y., Chen, X., ... & Long, X. (2025). CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interactive Geometry Refiner. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 5307-5317).
Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., ... & Guo, C. (2025). Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details. arXiv preprint arXiv:2506.16504.
Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., ... & Han, X. (2025). Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236, 3, 2.
Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., ... & Yang, J. (2025). Structured 3d latents for scalable and versatile 3d generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 21469-21480).
Huang, Z., Guo, Y. C., An, X., Yang, Y., Li, Y., Zou, Z. X., ... & Sheng, L. (2025). Midi: Multi-instance diffusion for single image to 3d scene generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 23646-23657).

Generation models: InstantMesh (Baseline)

Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., & Shan, Y. (2024). Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191.

Figure:
Line1: input image
Line2: InstantMesh result

Generation models: CraftsMan

Li, W., Liu, J., Yan, H., Chen, R., Liang, Y., Chen, X., ... & Long, X. (2025). CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interactive Geometry Refiner. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 5307-5317).

Figure:
Line1: input image
Line2: CraftsMan result

Generation models: Hunyuan3D 2.5

Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., ... & Guo, C. (2025). Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details. arXiv preprint arXiv:2506.16504.

Figure:
Line1: input image
Line2: Hunyuan3D result

Generation models: Hi3DGen

Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., ... & Han, X. (2025). Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236, 3, 2.

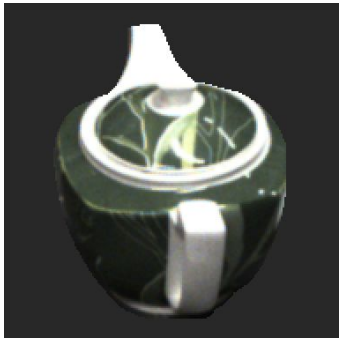
Figure:
Line1: input image
Line2: Hi3DGen result

Generation models: TRELLIS

Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., ... & Yang, J. (2025). Structured 3d latents for scalable and versatile 3d generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 21469-21480).







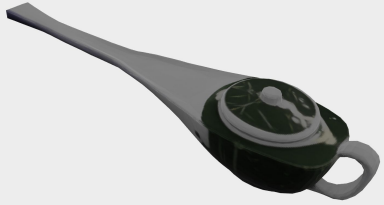



Figure:
Line1: input image
Line2: TRELLIS result

Generation models: MIDI-3D

Huang, Z., Guo, Y. C., An, X., Yang, Y., Li, Y., Zou, Z. X., ... & Sheng, L. (2025). Midi: Multi-instance diffusion for single image to 3d scene generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 23646-23657).

Figure:
Line1: input image
Line2: MIDI-3D result

Evaluation

Evaluation: overview

VGGT

- Result

Mesh

- a.method
- b.Target selection
- c.Result

Pose

- a.method
- b.Result

Evaluation: VGGT Depth

	Default		Normalized	
Scenes	MAE	RMSE	MAE	RMSE
scene01	0.4365651906	0.453864485	0.1134399325	0.2156497538
scene02	0.5740477443	0.9068885446	0.4324728251	1.090735674
scene03	0.6091852784	0.8730607629	0.333227694	0.5628105402
scene04	0.1677428633	0.334934473	0.162675187	0.3780018091
scene05	0.4045127034	0.4314068258	0.1791048348	0.3665329516
scene06	0.2744491696	0.2942522764	0.1165350527	0.2132981718
scene07	0.1806946397	0.3457638919	0.1686345637	0.4000940621
scene08	0.2258120626	0.2974425554	0.2116842866	0.3844485283
scene09	0.2883000076	0.3390727043	0.06309592724	0.1767561585
scene10	0.181447193	0.2007877678	0.04620821774	0.121305868
Mean	0.3342756853	0.4477474287	0.1827078521	0.3909633517

*Table. **MAE** and **RMSE** errors of estimated depths comparing to the ground truth*

MAE – Mean Absolute Error

RMSE – Root Mean Squared Error

Evaluation: Method

Align

The generated mesh would have different pose than groundtruth mesh

and the evaluation make sense only when these meshes are in the same pose

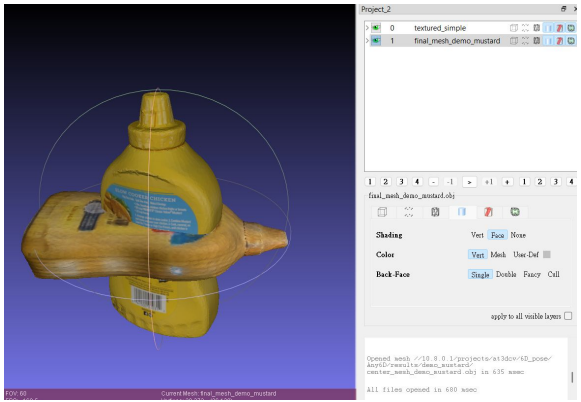


Figure . Meshed before alignment, horizontal mesh is generated mesh, vertical mesh is groundtruth mesh

Evaluation: Method

Align

We can end up have a aligned mesh for reconstruction evaluation

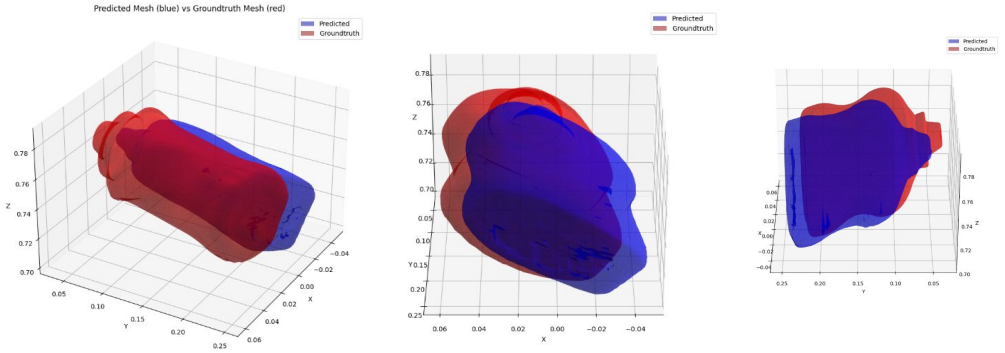
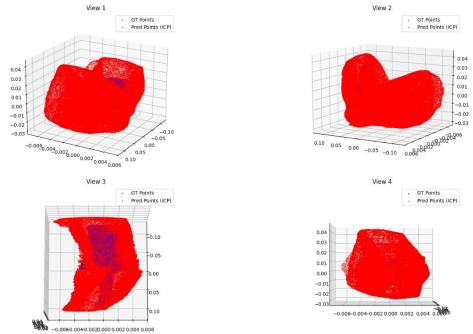


Figure . aligned mesh

Evaluation: Method

Chamfer distance

chamfer distance is a metric to compare the distance between two pointclouds, the lower the value is, the recosntruction performance is better



chamfer_distance = 0.56

Figure . chamfer_distance example

Evaluation: Target selection

Housecat6D

In Housecat6D, we choose the 000000.png in scene01-scene10

Figure . 000000.png in scene01

Jung, HyunJun and Wu, Shun-Cheng and Ruhkamp, Patrick and Zhai, Guangyao and Schieber, Hannah and Rizzoli, Giulia and Wang, Pengyuan and Zhao, Hongcheng and Garattoni, Lorenzo and Meier, Sven and Roth, Daniel and Navab, Nassir and Busam, Benjamin. HouseCat6D-A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios (CVPR2024)

Evaluation: Target selection

Baseline experiment

This is the chamfer distance for each scene and object by using the 3D generation method in Any6D

object chamfer distance

Object_Group	Average_Chamfer_Distance
remote	0.6763136182
cup	0.7215416
teapot	0.9388625208
can	1.051433736
bottle	1.211294348
shoe	1.320448426
cutlery	2.130128087
box	3.715149431
glass	6.061597882

Table. average chamfer distance in each object

scene chamfer distance

Scene_Group	Average_Chamfer_Distance
scene1	1.253926
scene2	1.675768
scene3	0.758823
scene4	1.222022
scene5	1.972883
scene6	0.596816
scene7	1.366534
scene8	1.260928
scene9	2.56075
scene10	3.627176

Table. average chamfer distance in each scene

Evaluation: Target selection

Baseline experiment

Low chamfer distance example

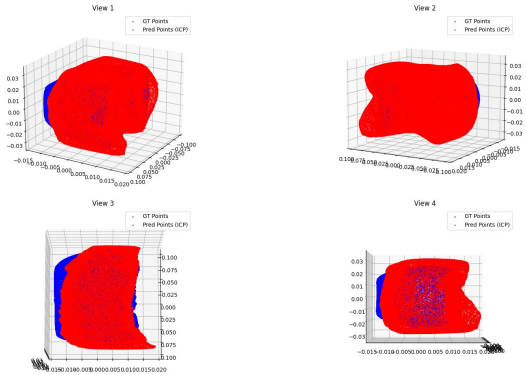


Figure.nicely generated remote from scene3

Evaluation: Target selection

Baseline experiment

High chamfer distance example

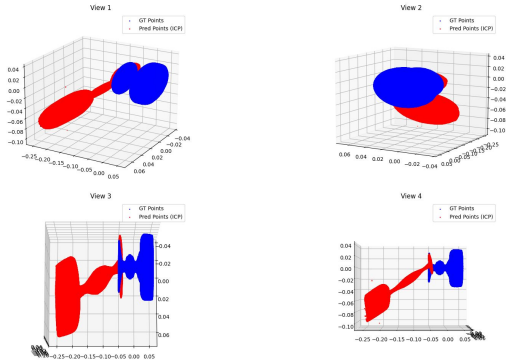
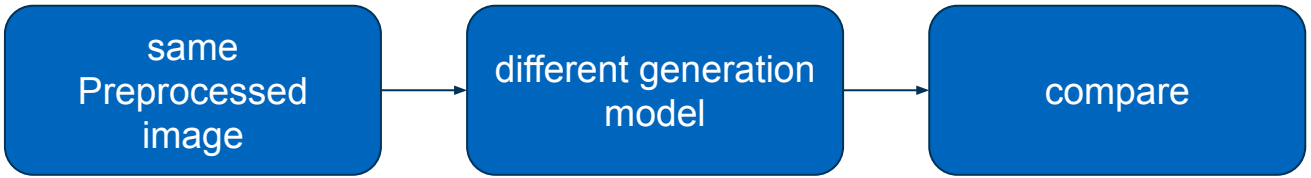


Figure.poorly generated glass from scene5

Evaluation: 3D generation model

Experiment on different models



Evaluation: 3D generation model

Experiment on different models

groundtruth		instantmesh	huanyuan	craftmen	trellis	midi-3d	hi3d-gen
scene01	can-kidney_beans	2.4438307	1.502351	2.389527	1.7705	2.338262	-
scene02	remote-black	0.8876515	1.557944	0.921144	1.545478	1.671484	2.791037
scene04	bottle-evian_frozen	11.591091	11.05339	12.49337	11.9053	17.52694	11.00451
scene04	box-barilla	1.0778612	1.166834	0.865721	0.610993	0.710358	1.411842
scene04	cup-white_whisker	6.6477307	8.693179	7.240895	7.571412	0.831209	0.178702
scene04	cutlery-spoon_1	0.2811144	0.353865	0.248515	0.3346	0.246362	0.2342312
scene05	glass-small	8.9108204	6.082433	10.67561	2.248441	5.677378	4.195016
scene05	remote-silver	10.485339	0.442109	10.20728	0.583899	5.682212	10.91002
scene06	cutlery-knife_1	0.5592384	18.08344	8.910761	4.790768	16.86956	8.317448
scene07	teapot-blue_floral	2.4636603	1.522847	3.204081	0.98607	1.101564	2.515233
scene08	bottle-85_alcool	19.32806	20.69918	21.81716	2.692343	19.35918	30.75065
scene08	shoe-white_viva_sandal_right	21.317511	1.464377	1.489781	11.67677	10.54955	1.540491
scene09	bottle-deodorant_spray	0.497184	1.750192	0.667985	0.407317	0.614657	0.554584
scene10	box-iglo	15.080177	12.255	2.535156	13.91971	14.30692	17.93114
scene10	teapot-green_grass	2.2419471	1.94699	1.981411	2.307847	2.111526	2.434425
	average	6.920881113	5.904942067	5.709893133	4.223429867	6.6398108	6.7692378

Table . Comparison between generation methods

Xu, Jiale and Cheng, Weihao and Gao, Yiming and Wang, Xintao and Gao, Shenghua and Shan, Ying. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models,arXiv preprint arXiv:2404.07191,2024

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Shaoxiong Yang, Song Zhang, Yang Liu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng, Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei Huang, Chunchao Guo. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation, arXiv:2501.12202

Xiang, Jianfeng and Lv, Zelong and Xu, Sicheng and Deng, Yu and Wang, Ruicheng and Zhang, Bowen and Chen, Dong and Tong, Xin and Yang, Jiaolong.Structured 3D Latents for Scalable and Versatile 3D Generation,arXiv preprint arXiv:2412.01506

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, Xiaoxiao Long. CraftsMan3D: High-fidelity Mesh Generation with 3D Native Generation and Interactive Geometry Refiner, arXiv:2405.14979

Huang, Zehuan and Guo, Yuan-Chen and An, Xingqiao and Yang, Yunhan and Li, Yangguang and Zou, Zi-Xin and Liang, Ding and Liu, Xihui and Cao, Yan-Pei and Sheng, Lu.Midi: Multi-instance diffusion for single image to 3d scene generation.Proceedings of the Computer Vision and Pattern Recognition Conference.2025

Ye, Chongjie and Wu, Yushuang and Lu, Ziteng and Chang, Jiahao and Guo, Xiaoyang and Zhou, Jiaqing and Zhao, Hao and Han, Xiaoguang.Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging.arXiv preprint arXiv:2503.22236.2025

Evaluation: 3D generation model

Experiment on different preprocessed image

Table . Comparison between preprocessed image

Evaluation: 3D generation model

Experiment on different preprocessed image

huanyuan		groundtruth	groundtruth_rembg	grounded_sam2_rembg
scene01	can-kidney_beans	1.502351	1.422792	1.354992
scene02	remote-black	1.557944	1.860731	1.862158
scene04	bottle-evian_frozen	11.05339	9.951886	10.11988
scene04	box-barilla	1.166834	0.757795	0.757533
scene04	cup-white_whisker	8.693179	1.230549	7.925604
scene04	cutlery-spoon_1	0.353865	0.327254	0.400229
scene05	glass-small	6.082433	11.34969	11.5542
scene05	remote-silver	0.442109	0.378744	0.378001
scene06	cutlery-knife_1	18.08344	5.070479	5.051104
scene07	teapot-blue_floral	1.522847	1.500943	1.531083
scene08	bottle-85_alcool	20.69918	17.69651	18.18188
scene08	shoe-white_viva_sandal_right	1.464377	1.982968	1.895802
scene09	bottle-deodorant_spray	1.750192	0.812461	0.84477
scene10	box-iglo	12.255	15.9899	15.58182
scene10	teapot-green_grass	1.94699	2.660366	2.410698
	average	5.904942067	4.866204533	5.323316933

Table . Comparison between preprocessed image

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Shaoxiong Yang, Song Zhang, Yang Liu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng, Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei Huang, Chunchao Guo. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation, arXiv:2501.12202

Evaluation: 3D generation model

Experiment on different preprocessed image

trellis		groundtruth	groundtruth_rembg	grounded_sam2_rembg
scene01	can-kidney_beans	1.7705	1.861531	1.123321
scene02	remote-black	1.545478	1.331444	0.947796
scene04	bottle-evian_frozen	11.9053	3.613688	5.236803
scene04	box-barilla	0.610993	1.215968	1.021512
scene04	cup-white_whisker	7.571412	7.67452	7.562767
scene04	cutlery-spoon_1	0.3346	0.241217	0.264786
scene05	glass-small	2.248441	4.510146	4.143532
scene05	remote-silver	0.583899	0.593474	0.829972
scene06	cutlery-knife_1	4.790768	4.703701	0.668507
scene07	teapot-blue_floral	0.98607	0.862542	1.522332
scene08	bottle-85_alcool	2.692343	6.092463	3.016116
scene08	shoe-white_viva_sandal_right	11.67677	1.855909	1.269793
scene09	bottle-deodorant_spray	0.407317	0.416816	3.537048
scene10	box-iglo	13.91971	15.85328	11.71336
scene10	teapot-green_grass	2.307847	2.367948	2.846897
	average	4.223429867	3.5463098	3.046969467

Table . Comparison between preprocessed image

Xiang, Jianfeng and Lv, Zelong and Xu, Sicheng and Deng, Yu and Wang, Ruicheng and Zhang, Bowen and Chen, Dong and Tong, Xin and Yang, Jiaolong.Structured 3D Latents for Scalable and Versatile 3D Generation,arXiv preprint arXiv:2412.01506

Evaluation: Pose evaluation

What is relative pose

Figure.view anchor image

Figure.view target image

Evaluation: Pose evaluation

ADD,ADD-S,R_error,T_error

ADD (Average Distance of Model Points)

- Calculates the mean distance between each point on a 3D model transformed by the estimated pose and the **corresponding** point transformed by the ground truth pose.

ADD-S (Average Distance of Model Points - Symmetric)

- Designed for symmetric or partially symmetric objects.
- For each model point under the estimated pose, ADD-S measures the distance to the ***closest*** point on the model under the ground truth pose, using a nearest neighbor search.

R_error,T_error

- R_error is the angular difference for predicted pose, measurement in **degrees** here
- T_error is the euclidean distance for predicted pose,measuremen in **centimeter** here

Evaluation: Pose evaluation

Scene	ADD-S	ADD	R_error	T_error
scene01	0.92	1.02	178.64	101.24
scene02	1.066667	1.11666	167.0167	112.4667
scene03	1.2	1.26666	179.5167	126.9667
scene04	1.566667	1.61666	179	161.9833
scene05	1	1.11666	153.5333	108.6667
scene06	1.033333	1.05	179.3167	105.1833
scene07	1.133333	1.2	152.1	118.7
scene08	1.416667	1.48333	168.45	147.9833
scene09	1.283333	1.33333	178.6333	134.05
scene10	1.45	1.53333	153.9667	152.1333
OVERALL	1.211864	1.27796	168.8542	127.3729

Table . Any6D pose evaluation result on housecat6d

Conclusion

- **Complete Pipeline Construction**
  We built a full pipeline from **targets in RGB images to 3D mesh reconstruction and finally to 6D pose estimation**.

- **Comprehensive Research at Each Stage**
  - Depth prediction
  - Extract the target from img and do mesh generation
  - From RGBD and 3D mesh to 6D pose estimation

Conclusion

- **Extensive Module Testing and Comparison**
  - Depth generation
    i. VGGT got good result for depth prediction
  - Mesh generation
    i. Non-texture mesh can has the same performance as texture mesh
    ii. The new end-to-end generation model outperforms the old methods
- **Limitations**
  - Fully test all modules and do comprehensive evaluation.
  - Current Any6D will give slightly different pose, which seems to be local minimum for the right pose.

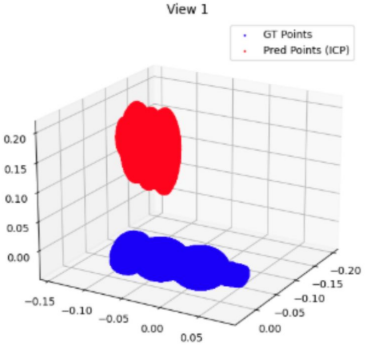
What’s Next?

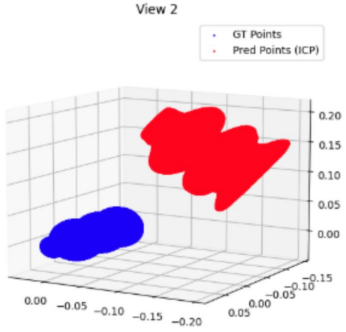
What's Next:

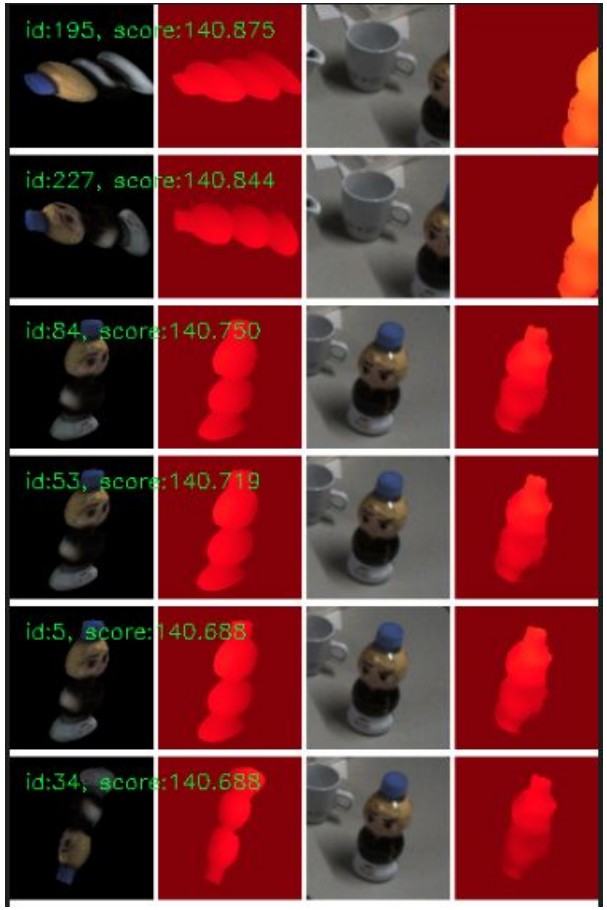
Current Problem:

The error rate of Any6D is too high. The output of Any6D is hard to control.

- Any6D uses a NN for scoring, but the score does not necessarily increase as the predicted pose gets closer to the ground truth.
- sometimes assigns the highest score to a clearly incorrect result.

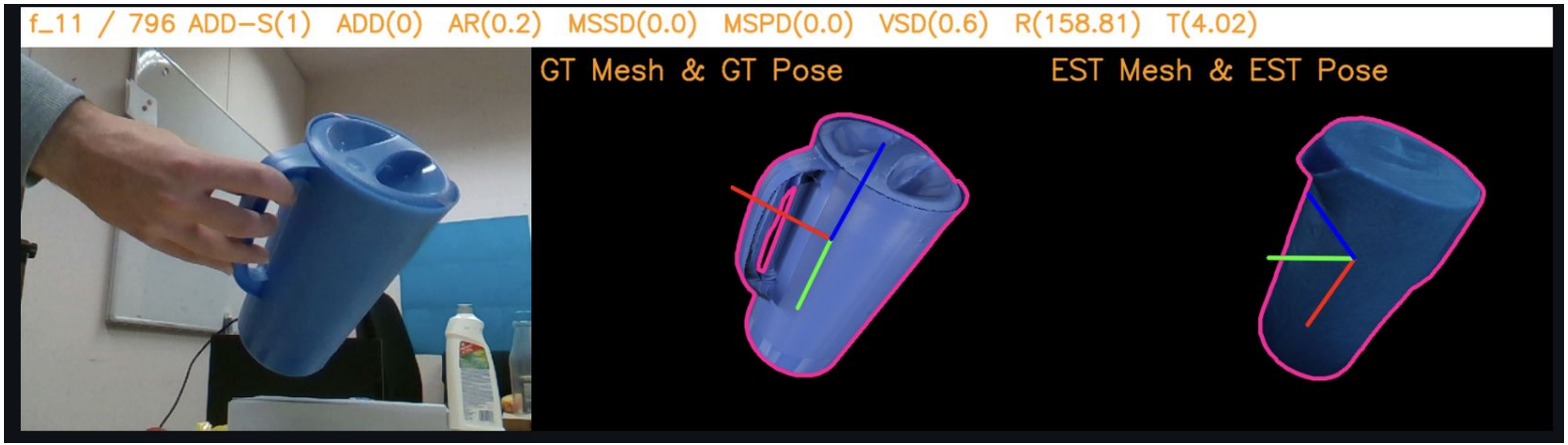






What's Next: Overview

The pose from Any6D is not as good as they claimed:



Research Plan:
get a better pose estimation

Thanks!

Appendix

Evaluation: Method

Align

In any6d, the same input for anchor and target image can help align the mesh and calculate the chamfer distance

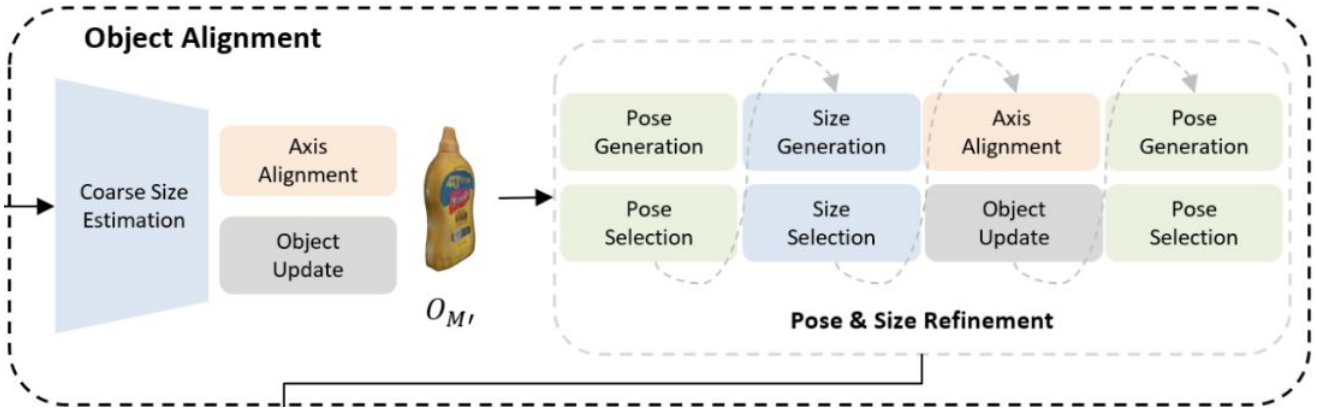


Figure . The alignment procedure in Any6D

Lee, Taeyeop and Wen, Bowen and Kang, Minjun and Kang, Gyuree and Kweon, In So and Yoon, Kuk-Jin. {Any6D}: Model-free 6D Pose Estimation of Novel Objects, (CVPR2025)

Evaluation: Method

Align

Several size and pose of mesh would be created and applied to generated mesh, then be scored with transformer.

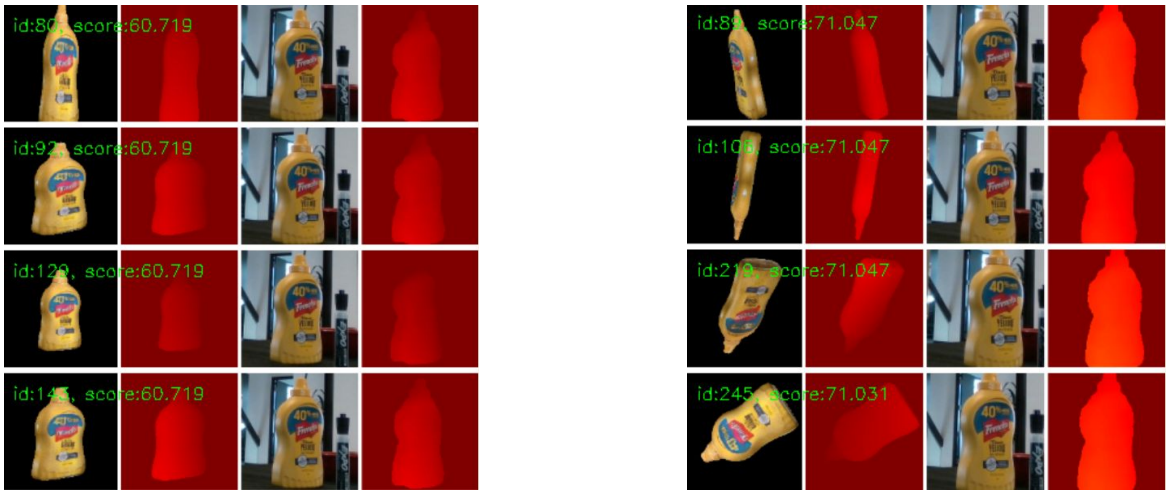


Figure . Left image is the score for different size, right image is score for different pose

Evaluation: Method

Align

We change the resized from non-uniform scale in xyz, to an uniform scale. So that we can have an idea how good the structure of generated mesh is

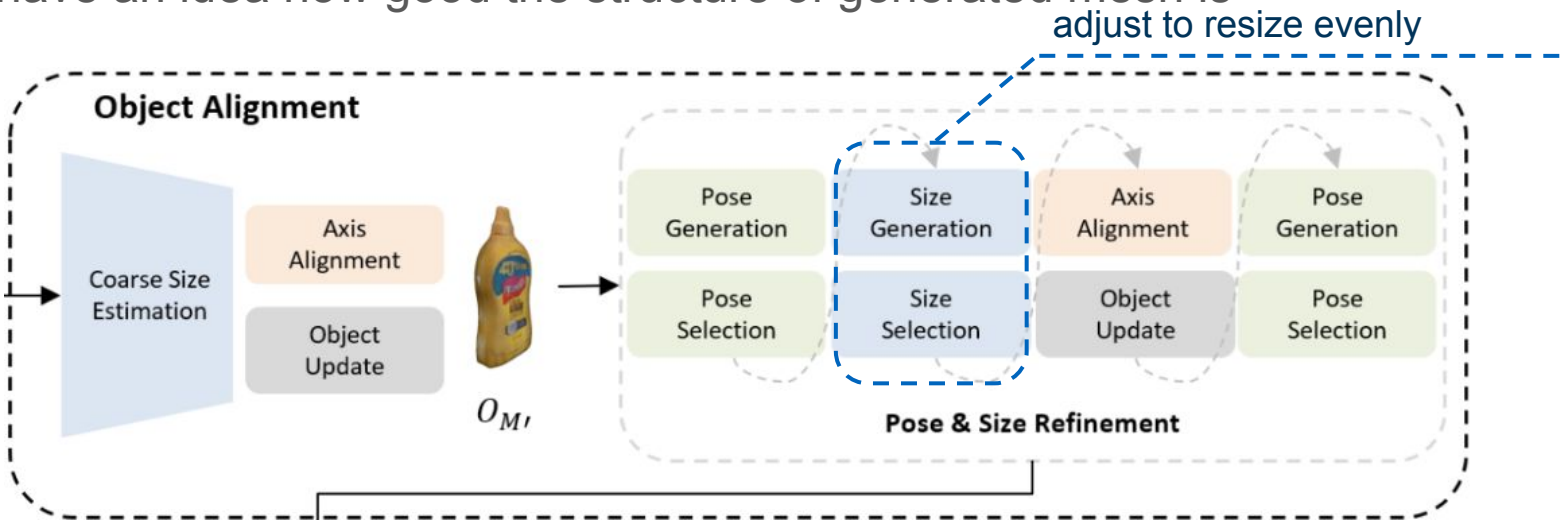


Figure . The adjusted even_resized in Any6D

Lee, Taeyeop and Wen, Bowen and Kang, Minjun and Kang, Gyuree and Kweon, In So and Yoon, Kuk-Jin. {Any6D}: Model-free 6D Pose Estimation of Novel Objects, (CVPR2025)