VGGT : Visual Geometry Grounded Transformer

🏆 *Best award*

Paper Review

2025.09.04

Changseon Yu

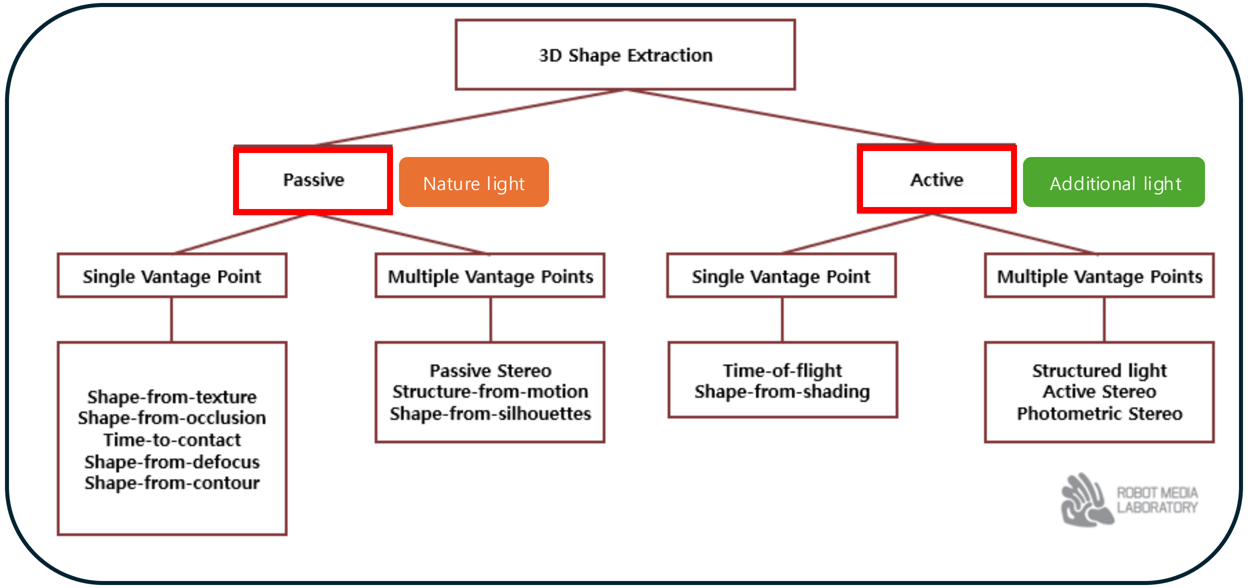
Contents

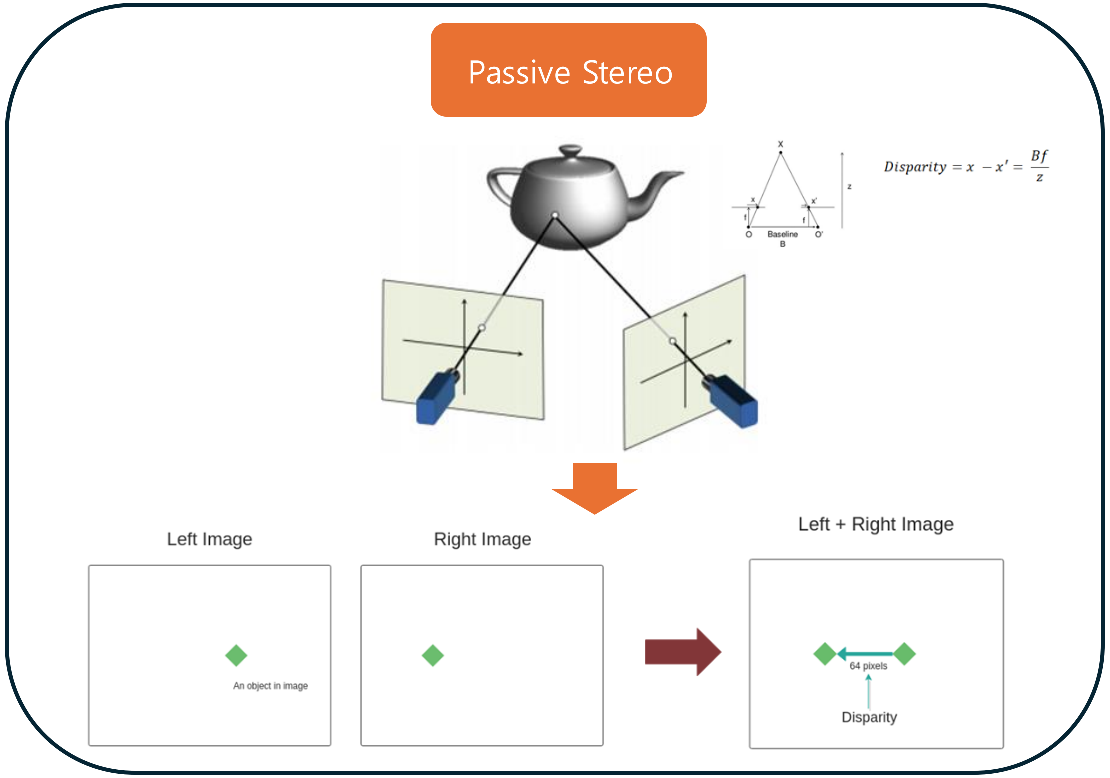
Introduction

Background(1)

❖ 3D reconstruction

- The process of creating a 3D model of a target object or scene using 2D images or other data (eg. Laser scan)
- Robotics, Virtual Reality (VR), Autonomous Driving



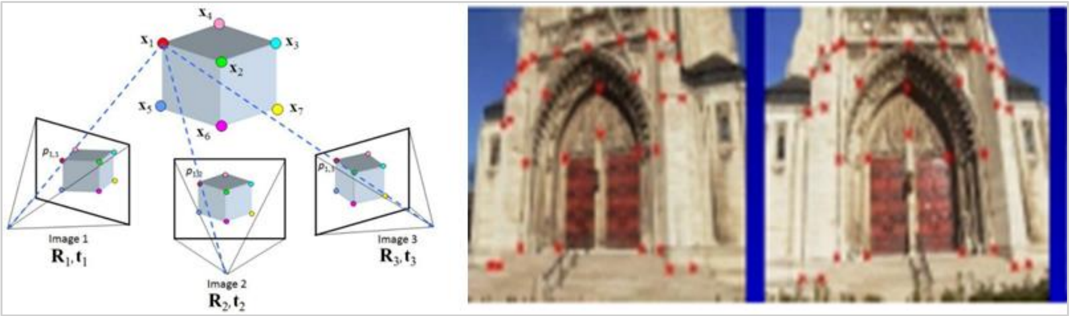


Introduction

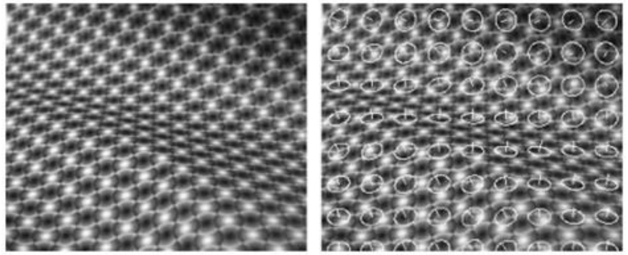
Background(1)

❖ **3D reconstruction – Passive**

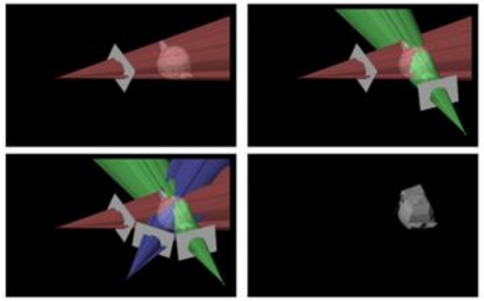
Structure-from-Motion(SFM)



Shape-from-Texture and Shape-from-Contour



Shape-from-Sillhouettes



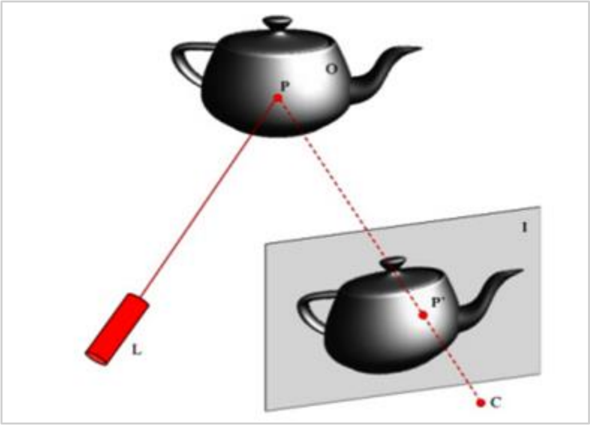
Introduction

Background(1)

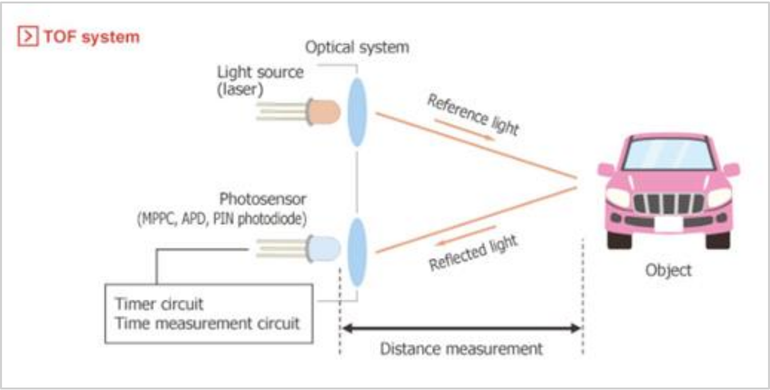
❖ 3D reconstruction – Active

- A technique of using additional light
- Difficult to collect data because of using other devices like lasers and make data
- Reliability is high due to low impact on surrounding environment

Active Stereo



Time-of-Flight(ToF)

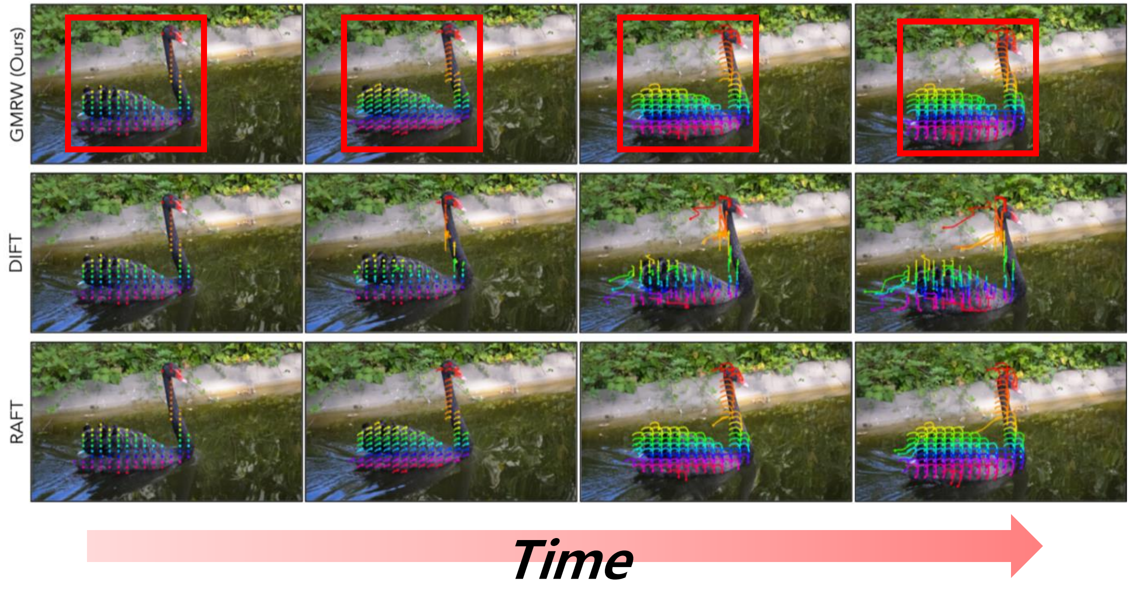


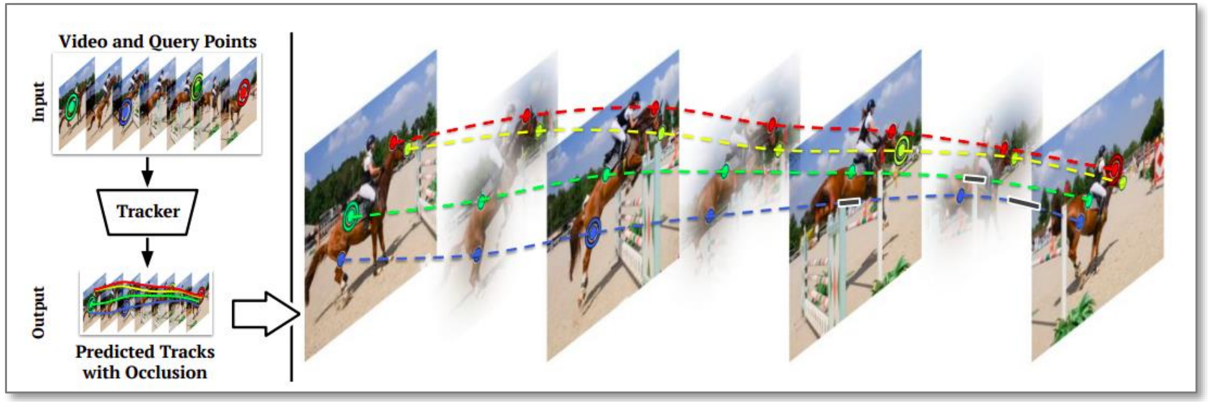
Introduction

Background(2)

❖ **Tracking-Any-Point**

- Track points of interest throughout the video sequence, including dynamic movements
- Predict the 2D positions corresponding to these points in all different frames (Input video & 2D query points)





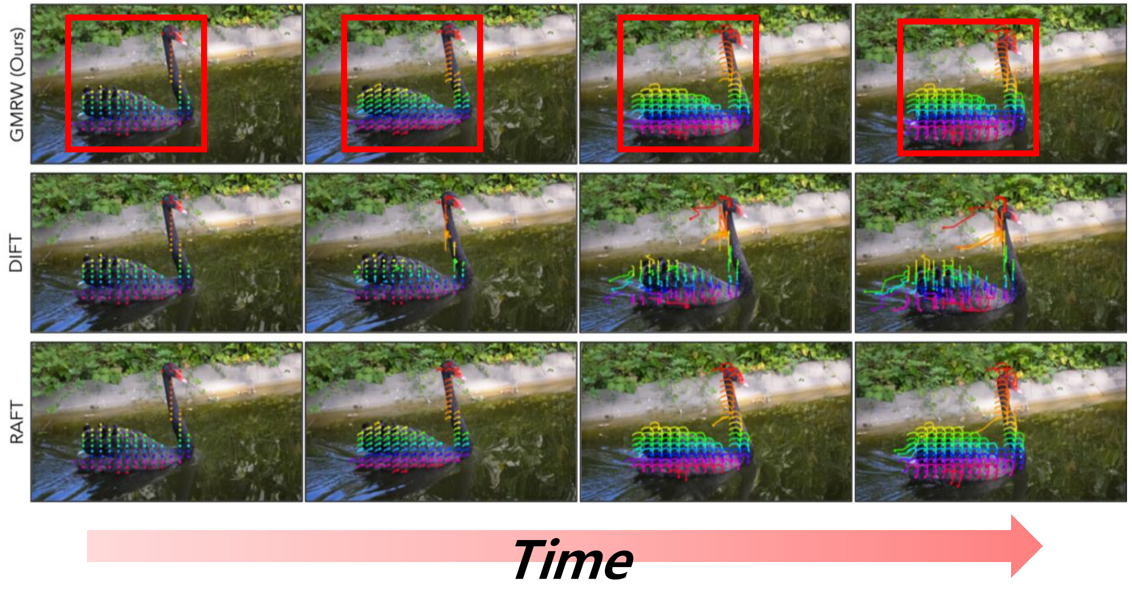
Shrivastava, Ayush, and Andrew Owens. "Self-Supervised Any-Point Tracking by Contrastive Random Walks." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
Doersch, Carl, et al. "Tap-vid: A benchmark for tracking any point in a video." Advances in Neural Information Processing Systems 35 (2022): 13610-13626.

Introduction

Background(2)

❖ **Tracking-Any-Point**

- Track points of interest throughout the video sequence, including dynamic movements
- Predict the 2D positions corresponding to these points in all different frames (Input video & 2D query points)

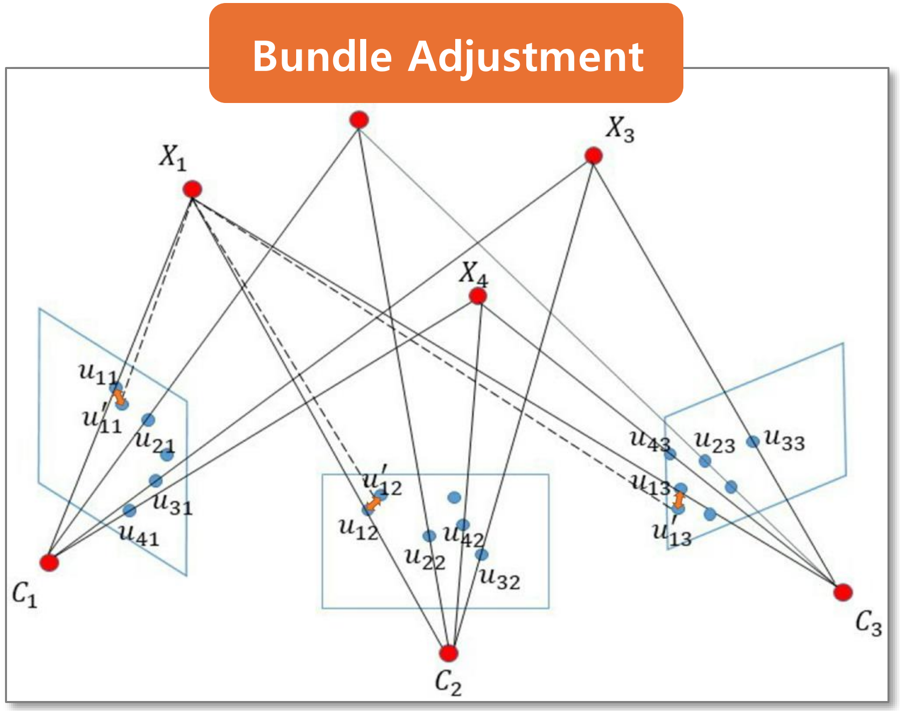


Shrivastava, Ayush, and Andrew Owens. "Self-Supervised Any-Point Tracking by Contrastive Random Walks." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
Doersch, Carl, et al. "Tap-vid: A benchmark for tracking any point in a video." Advances in Neural Information Processing Systems 35 (2022): 13610-13626.

Purpose

❖ Limitation of prior works

- Increasing complexity and computational cost of iterative optimization (eg. Bundle Adjustment)
- Processing images in pairs → Reconfigure multiple images with post-processing



Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Purpose

❖ Limitation of prior works

- All large scale 3D neural networks, but models each focused on a single 3D task
- Transformer based large scale foundation model

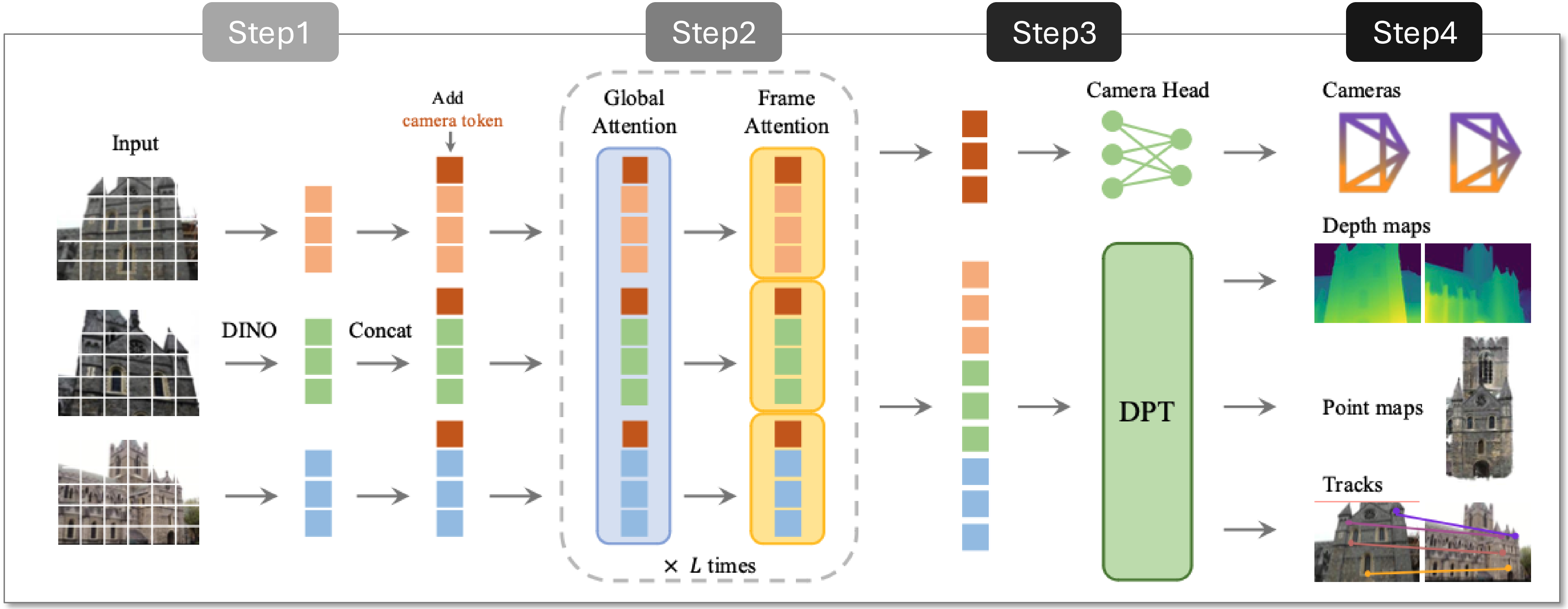
NeurIPS — Depth Anything →

CVPR — MoGe-2 → MoGe-2

ICLR — LRM →

Method

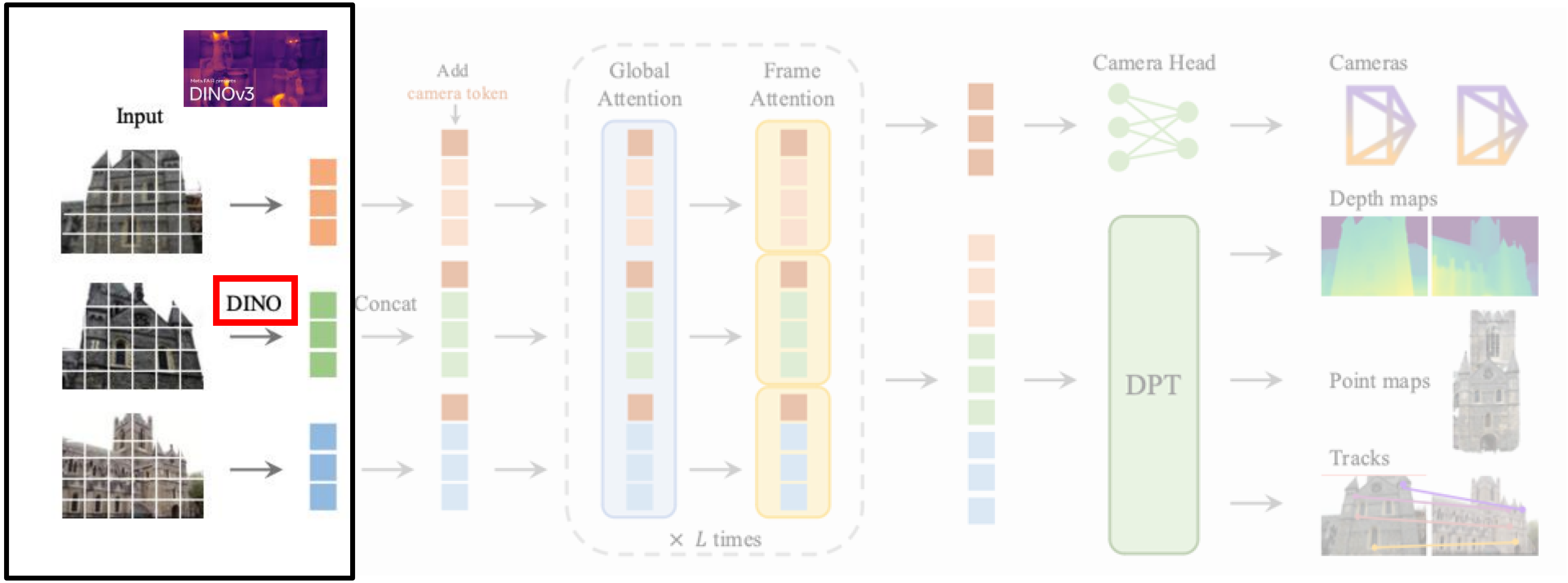
❖ Overall Architecture



Purpose

❖ **Step1. – Input**

- Input: N RGB image sequences as inputs (can be entered in any order)
- Input images are converted to tokens through Pretrained DINO
- Add a camera token to each image for camera parameter prediction

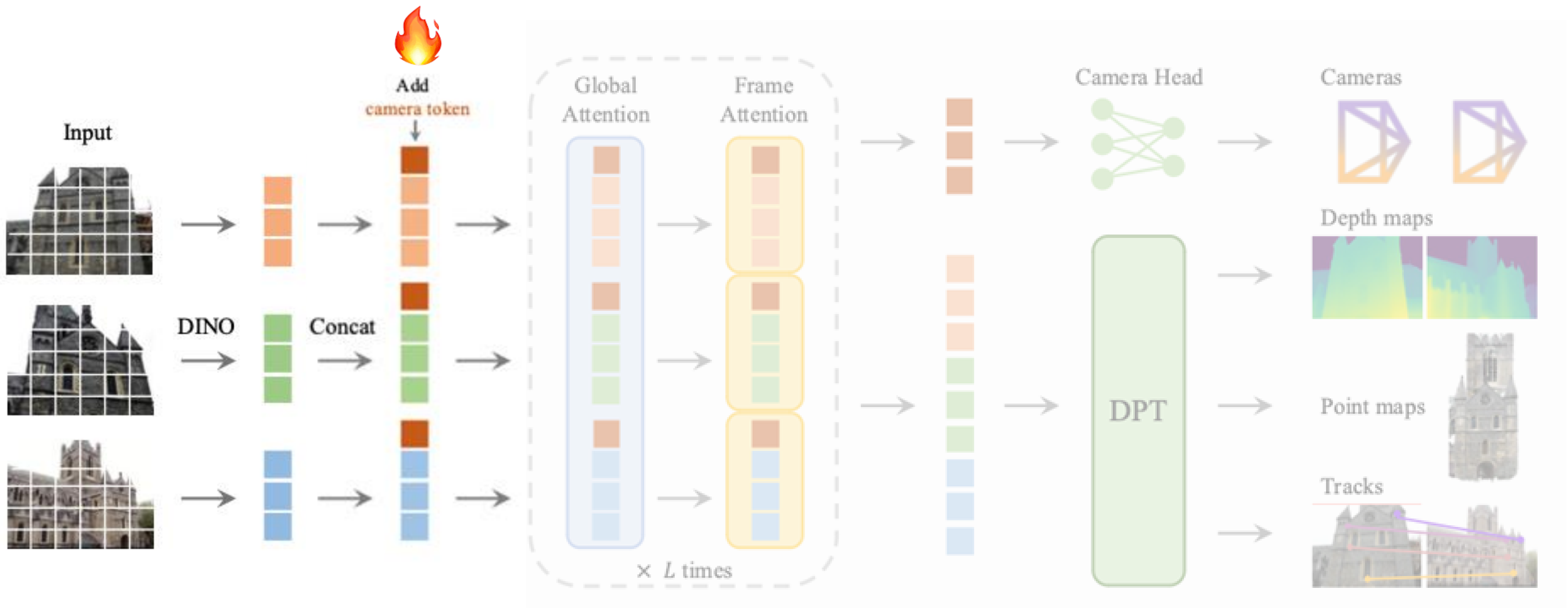


Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Purpose

❖ **Step1. – Input**

- Input: N RGB image sequences as inputs (can be entered in any order)
- Input images are converted to tokens through Pretrained DINO
- Add a camera token to each image for camera parameter prediction

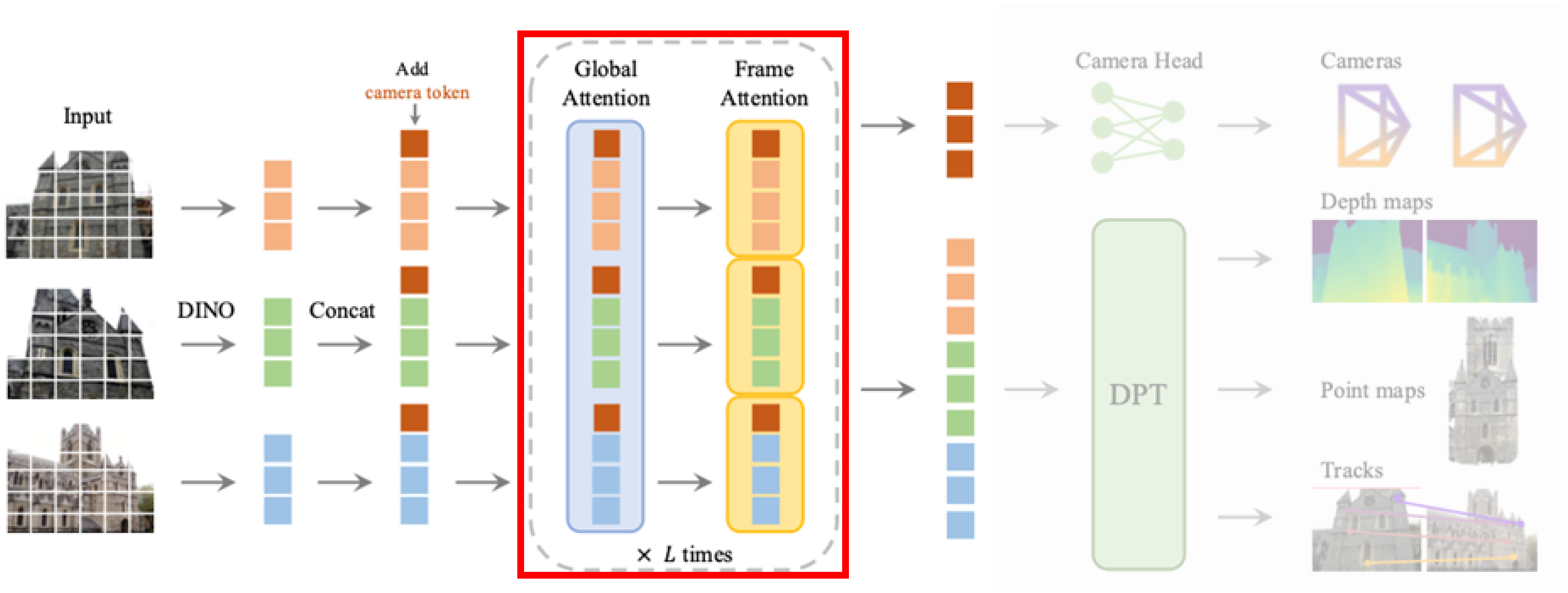


Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Purpose

❖ **Step2. – Alternating Attention(AA)**

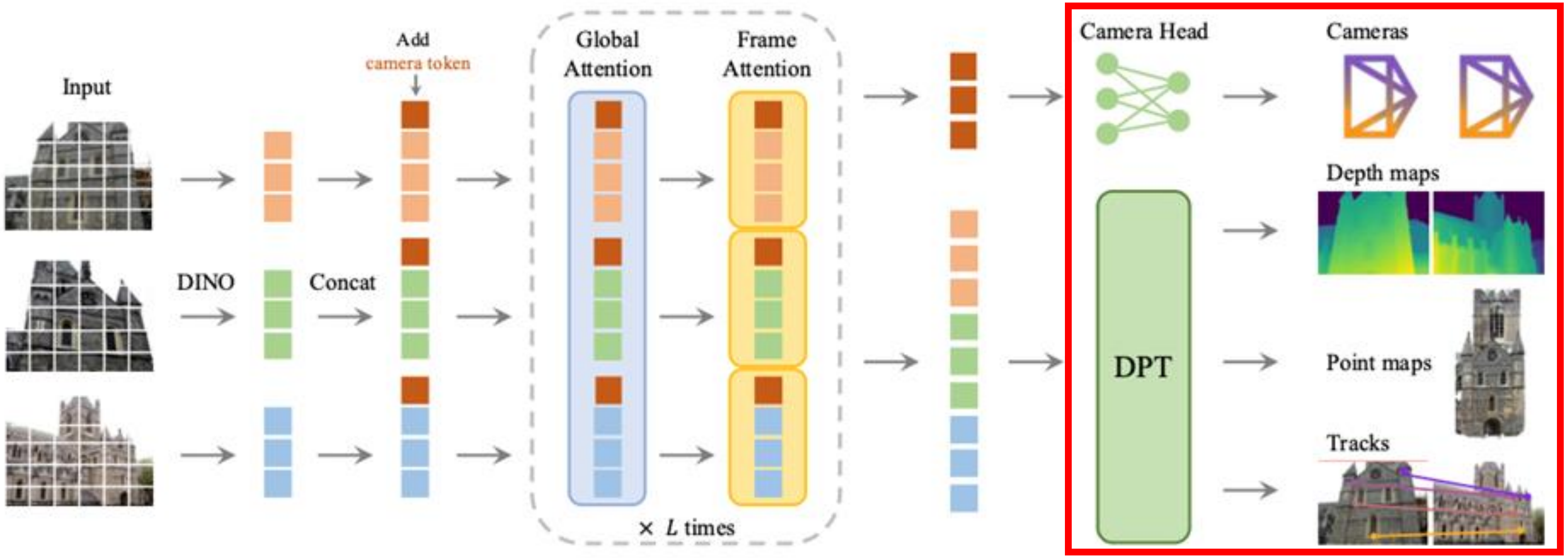
- Perform attention on all tokens in all frames
- Synthesizes information between images, and learn relationships between images



Purpose

❖ **Step3. – Prediction**

- Camera Head : Predict of the camera's internal/external parameters from the camera token
- DPT : Perform Dense Prediction, such as Depth Map, Point Map, etc. from image token



Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Purpose

❖ **Training Losses**

$$\mathcal{L} = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pmap}} + \lambda \mathcal{L}_{\text{track}}.$$

Camera Loss

$$\mathcal{L}_{\text{camera}} = \sum_{i=1}^{N} \left\| \hat{\mathbf{g}}_i - \mathbf{g}_i \right\|_{\epsilon}$$

Depth map Loss

$$\mathcal{L}_{\text{depth}} = \sum_{i=1}^{N} \|\Sigma_i^D \odot (\hat{D}_i - D_i)\| + \|\Sigma_i^D \odot (\nabla \hat{D}_i - \nabla D_i)\| - \alpha \log \Sigma_i^D$$

Point map loss

$$\mathcal{L}_{\text{pmap}} = \sum_{i=1}^{N} \|\Sigma_i^P \odot (\hat{P}_i - P_i)\| + \|\Sigma_i^P \odot (\nabla \hat{P}_i - \nabla P_i)\| - \alpha \log \Sigma_i^P$$

Tracking loss

$$\mathcal{L}_{\text{track}} = \sum_{j=1}^{M} \sum_{i=1}^{N} \|\mathbf{y}_{j,i} - \hat{\mathbf{y}}_{j,i}\|$$

Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Result

❖ Visualization of Point Map Estimation

- Point map visualization generated by the VGGT model
- VGGT works well in different kinds of scenes, such as buildings, indoor props, and outdoor environments
- Complex textures such as asphalt are also excellent

Result

❖ Camera Pose Estimation

- Red : Existing optimization-based method: longer than 10 seconds
- Blue : slightly run time improved model but still take a long time
- Green: Time has gotten a lot faster but AUC score is too low
- **Purple (Our model)** : Runtime is very fast. + Bundle adjustment slows down time but increases AUC socre

Methods	Re10K *(unseen)* AUC@30 ↑	CO3Dv2 AUC@30 ↑	Time
Colmap+SPSG [92]	45.2	25.3	∼ 15s
PixSfM [66]	49.4	30.1	> 20s
PoseDiff [124]	48.0	66.5	∼ 7s
DUSt3R [129]	67.7	76.7	∼ 7s
MASt3R [62]	76.4	81.8	∼ 9s
VGGSfM v2 [125]	78.9	83.4	∼ 10s
MV-DUSt3R [111] ‡	71.3	69.5	∼ 0.6s
CUT3R [127] ‡	75.3	82.8	∼ 0.6s
FLARE [156] ‡	78.8	83.3	∼ 0.5s
Fast3R [141] ‡	72.7	82.5	∼ **0.2s**
Ours (Feed-Forward)	85.3	88.2	∼ **0.2s**
Ours (with BA)	**93.5**	**91.8**	∼ 1.8s

Result

❖ Dense MVS Estimation

- 3D reconstruction task using multi view stereo images
- Competitive in performance with the latest MVS methods, even if you don't know the actual camera parameter.

Known GT camera	Method	Acc.↓	Comp.↓	Overall↓
✓	Gipuma [40]	**0.283**	0.873	0.578
✓	MVSNet [144]	0.396	0.527	0.462
✓	CIDER [139]	0.417	0.437	0.427
✓	PatchmatchNet [121]	0.427	0.377	0.417
✓	MASt3R [62]	0.403	0.344	0.374
✓	GeoMVSNet [157]	0.331	**0.259**	**0.295**
✗	DUSt3R [129]	2.677	0.805	1.741
✗	Ours	**0.389**	**0.374**	**0.382**

Table 2. **Dense MVS Estimation on the DTU [51] Dataset.** Methods operating with known ground-truth camera are in the top part of the table, while the bottom part contains the methods that do not know the ground-truth camera.

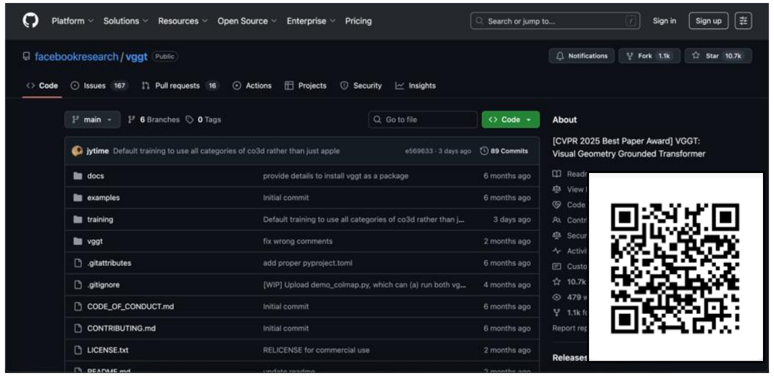
Result

❖ Visualization of Rigid and Dynamic Point Tracking

- No matter input images sequence
- Dynamic: Follow the Keypoint's trajectory

Figure 5. **Visualization of Rigid and Dynamic Point Tracking.** Top: VGGT's tracking module $\mathcal{T}$ outputs keypoint tracks for an unordered set of input images depicting a static scene. Bottom: We finetune the backbone of VGGT to enhance a dynamic point tracker CoTracker [56], which processes sequential inputs.

Conclusion



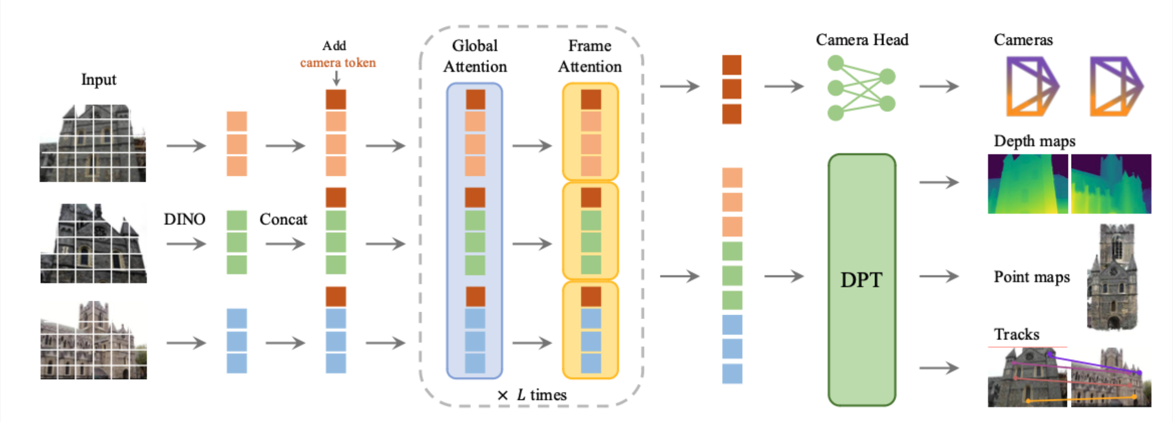
https://github.com/facebookresearch/vggt

❖ Limitation

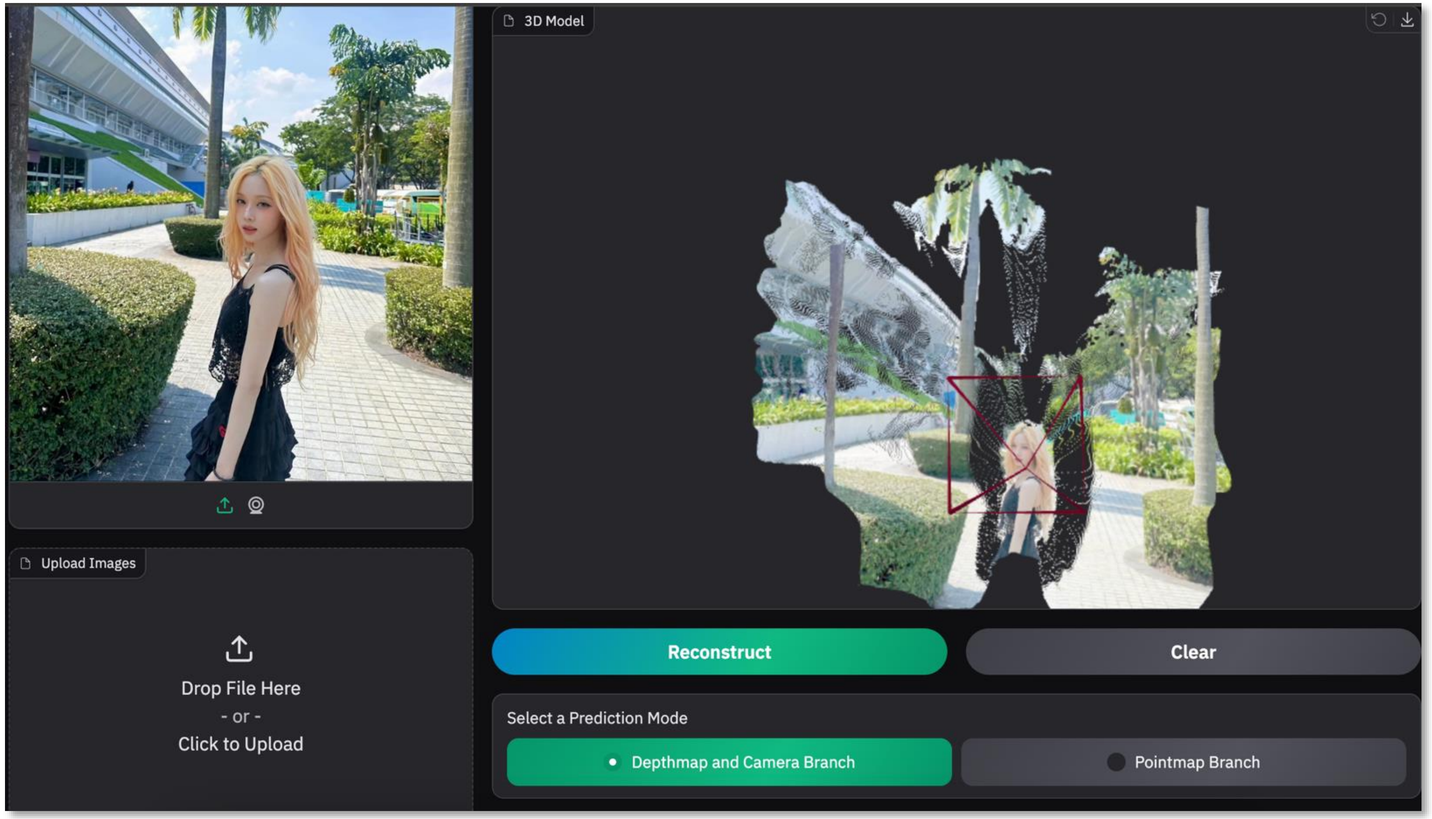
- Does not support fisheye or panoramic images
- Performance degrades at extreme input rotation
- It can handle minor non-rigid body movements, but fails in scenes with large variations

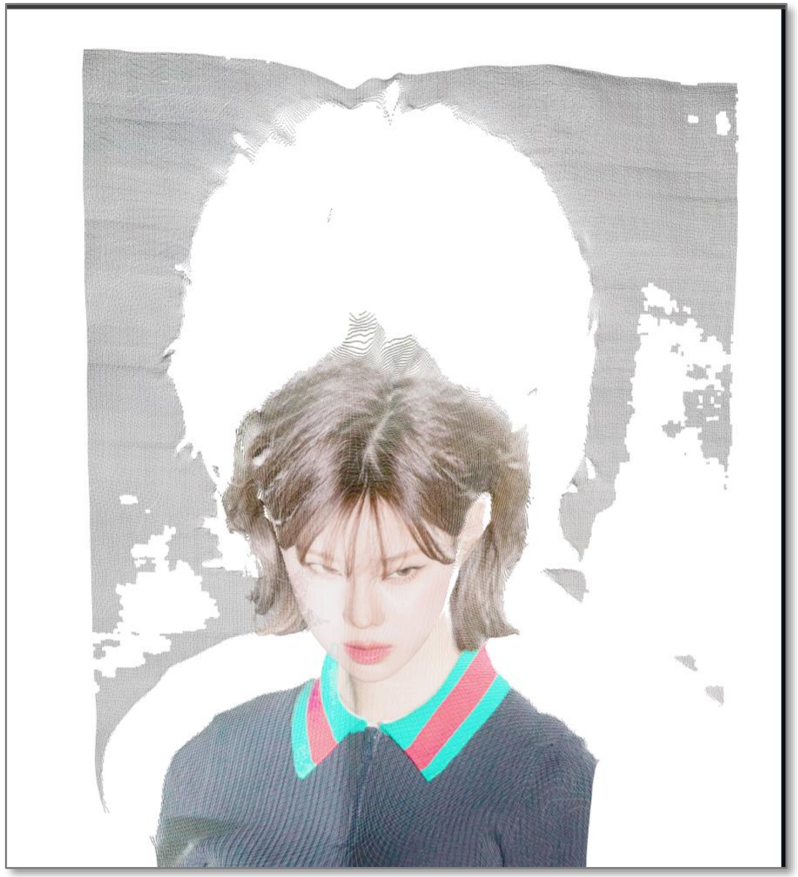
❖ Conclusion

- A feedforward neural network that can directly estimate all key 3D scene properties for hundreds of input views
- State-of-the-art performance in camera parameter estimation, multi-view depth estimation, dense point cloud reconstruction, and 3d point tracking



Code


3D Model
Upload Images
Drop File Here
- or -
Click to Upload
Reconstruct
Clear
Select a Prediction Mode
Depthmap and Camera Branch
Pointmap Branch



감사합니다.

End of Document

CHUNG-ANG UNIVERSITY