GigaVision Challenge

When Gigapixel Videography Meets Computer Vision

Track：Reconstruction

Team name：DTM3D

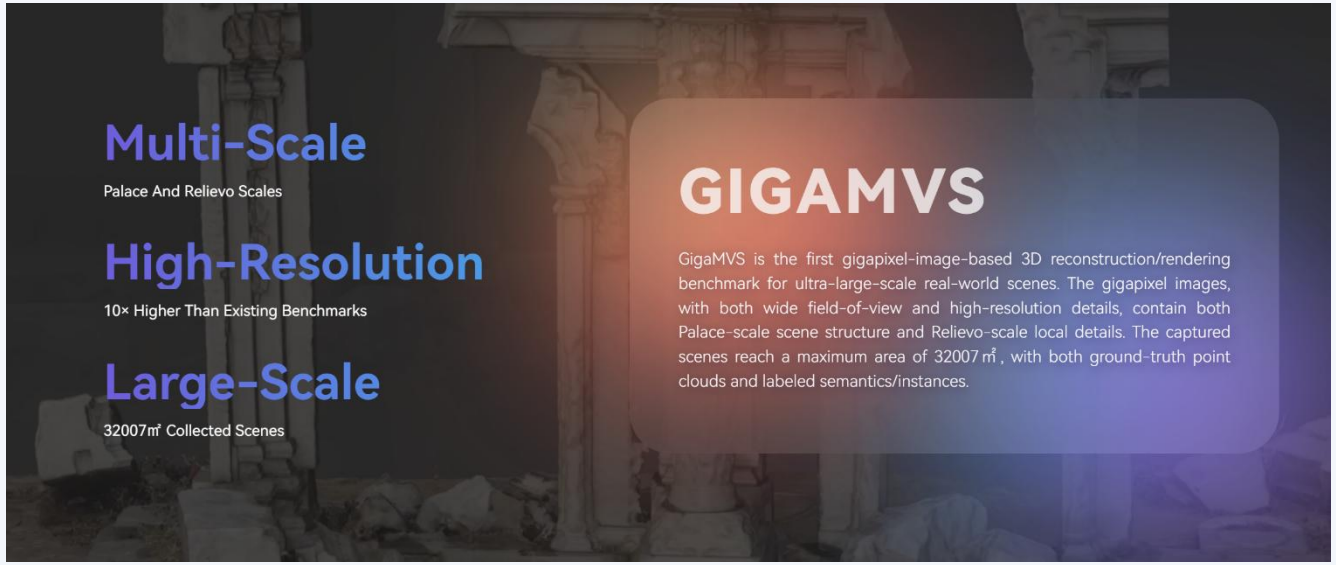
Team Introduction

Zizhuang Wei is currently an **AI algorithm researcher** in **Digital Twin Lab, Huawei**. He received the Ph.D degree from Graphics and Interaction Lab, Dept. of EECS, Peking University. His research interests focus on 3D reconstruction and deep learning.

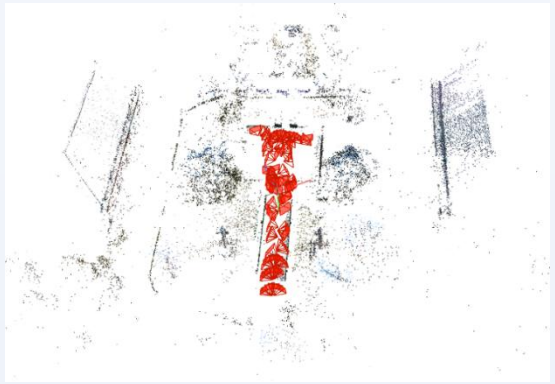
Qingtian Zhu is currently a **master student** at Graphics and Interaction Lab (GIL) of **Peking University**. His research interests include 3D reconstruction and computational photogrammetry.

GIGAVISION

Task Analysis



Original Images



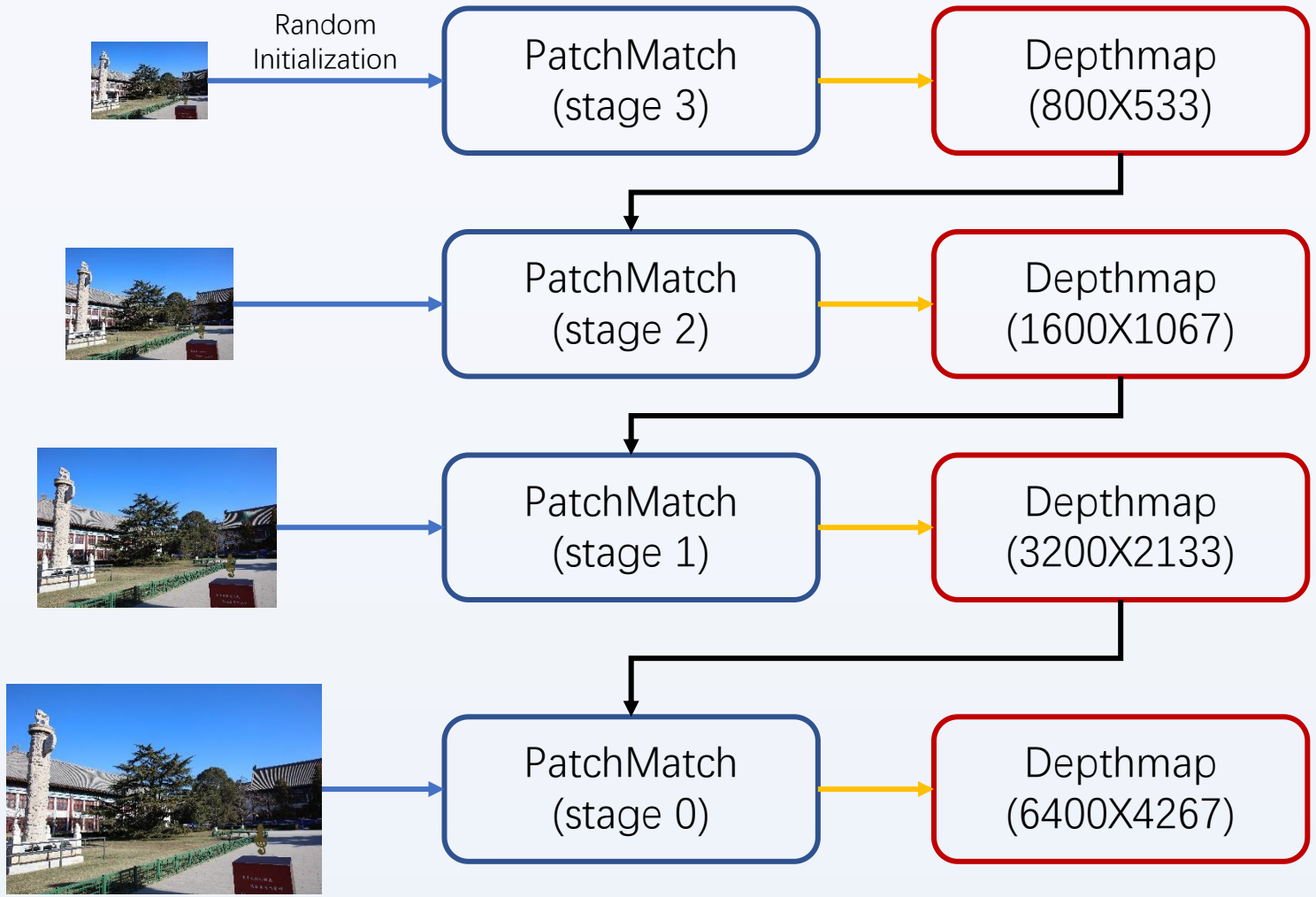
Camera poses

Challenges

- Very high resolution
- Weakly textured walls and floors
- Sparse view reconstruction
- Severe occlusion
- Unbounded scenario
- Large area of sky
- Complex lighting conditions
- ……

GIGAVISION

Solution and Innovation

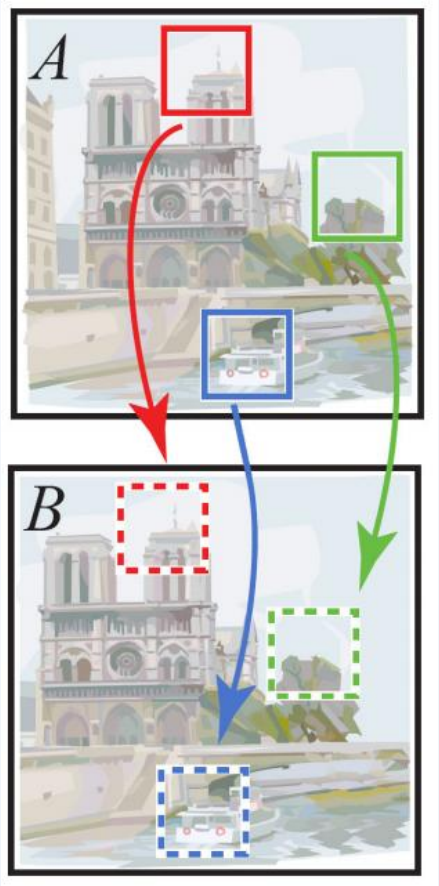


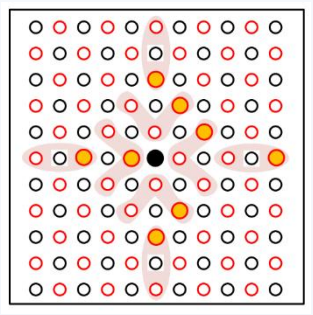
Multiscale depth inference

Multi-scale depth estimation framework is used to enhance weak texture regions, and **PatchMatch** method is used for feature matching at each stage (Learning features are not used due to memory limitations).

GIGAVISION

Solution and Innovation





$$\mathrm{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,N-1} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{8,1} & m_{8,2} & \cdots & m_{8,N-1} \end{bmatrix}$$

$$\rho_l^m = \frac{\mathrm{cov}_w(\boldsymbol{w}_l, \boldsymbol{w}_l^m)}{\sqrt{\mathrm{cov}_w(\boldsymbol{w}_l, \boldsymbol{w}_l)\ \mathrm{cov}_w(\boldsymbol{w}_l^m, \boldsymbol{w}_l^m)}}$$

Adaptive checkerboard sampling scheme[1] is used for parallel propagation on GPU.

Multi-Hypothesis joint view selection[1] is used to reduce the impact of bad views.

Bilaterally weighted NCC[2] is used to measure the multi-view similarity.

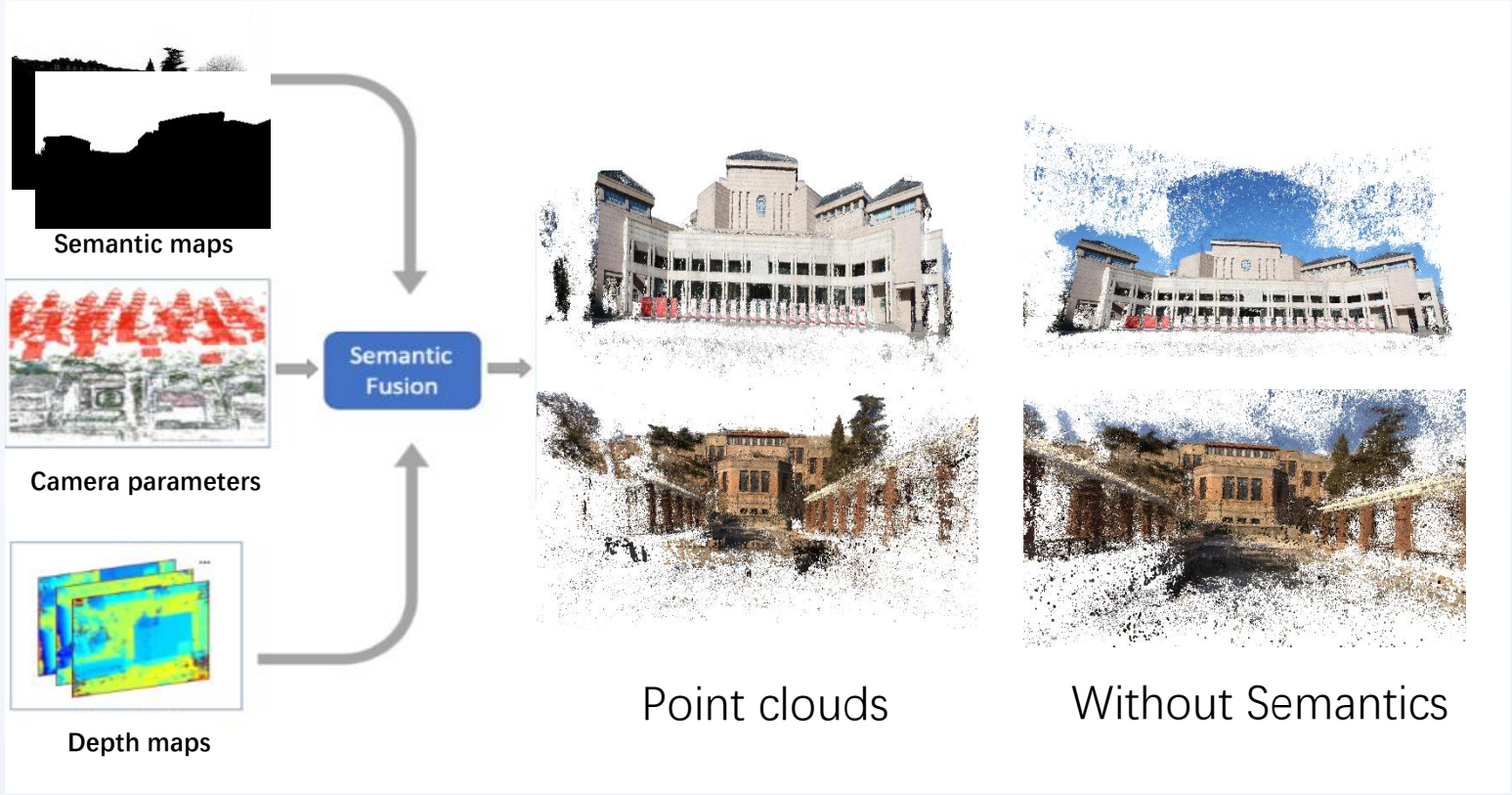
PatchMatch and propagation

[1] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5483–5492, 2019.

[2] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In European Conference on Computer Vision, pages 501–518. Springer, 2016

GIGAVISION

Solution and Innovation



Semantic & depth fusion

Dynamic depth map fusion algorithm[3] is used to filter the unreliable depths , while **Semantic maps** is used to remove the error points generated by sky area.

[3] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In European Conference on Computer Vision, pages 674–689. Springer, 2020.

GIGAVISION

Preparation

Camera parameters

```
extrinsic
E00 E01 E02 E03
E10 E11 E12 E13
E20 E21 E22 E23
E30 E31 E32 E33

intrinsic
K00 K01 K02
K10 K11 K12
K20 K21 K22
```

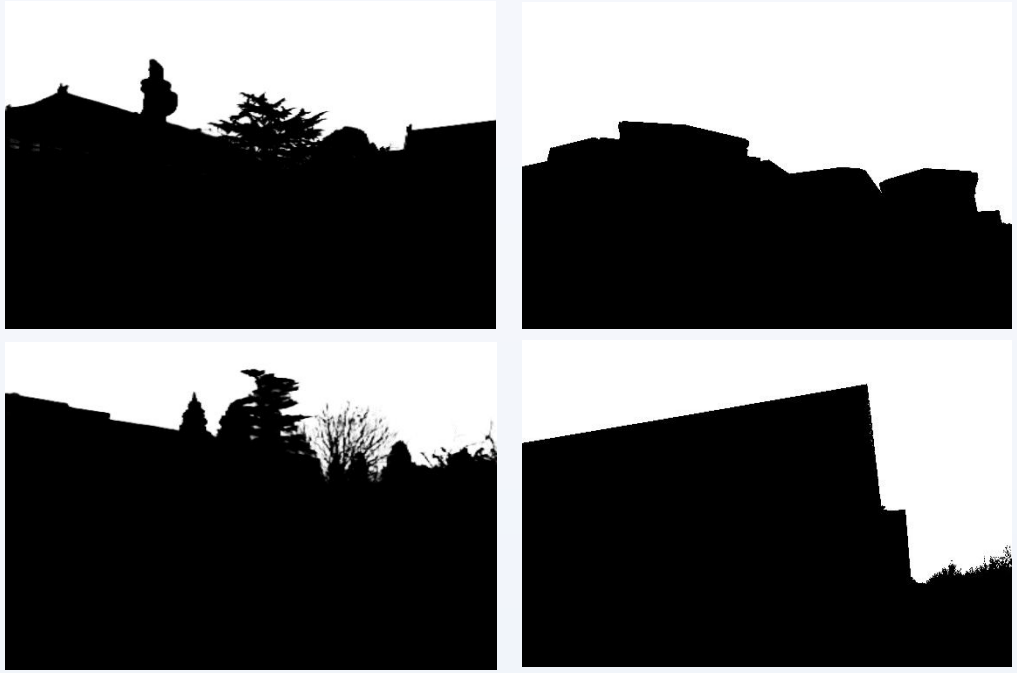
pair.txt

```
TOTAL_IMAGE_NUM
IMAGE_ID0
10 ID0 SCORE0 ID1 SCORE1 ...
IMAGE_ID1
10 ID0 SCORE0 ID1 SCORE1 ...
...
```

(depth range)
depth_min depth_inverval depth_num depth_max

Colmap SfM[4] is used to reconstruct the sparse point clouds. Then depth range and pair.txt are calculated by the sparse reconstruction for completing the camera files.

Semantics masks



DeepLab V3+[5] is used to segment images into **sky region and ground region**. We mask the sky region in white and the ground region in black.

[4] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4104–4113, 2016.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. TPAMI, 2017.

GIGAVISION

Preparation

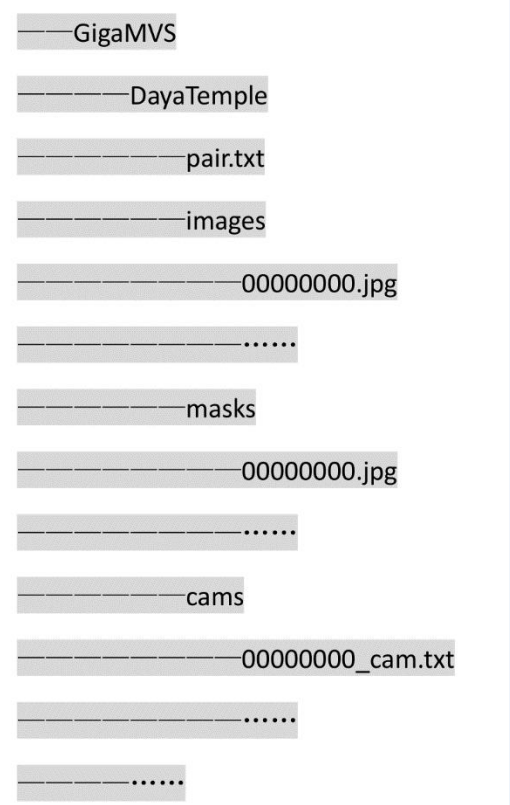
Requirements

Subject	Requirement
OS	Linux Ubuntu 20.04.5
GPU	24G Titan RTX at least
CPU	Intel Core I7+
Memory	256G+
Disk	4T
Cuda	>= 6.0
OpenCV	>=2.4
Cmake	3.25.1
Python	3.6

Settings

Subject	Requirement
Max resolution	6400
Patch size	21
Stages	4
View num	11
Iteration num	11
Hypotheses	8
Max view num (Fusion)	5
Depth threshold	0.03~0.09

File organization



It takes about one week to process all eight scenes.

Results

Leaderboard

#	Team	Members	Precision	Recall	F – score	Method	Code	Paper
1	ewrfcas		26.566373	22.35344	**22.33566**			
2	DTM3D		21.544648	**24.72914**	21.88166			
3	打赢baseline…		23.389852	19.22023	20.45153			
4	WeikangYou		22.684377	19.21984	19.99421			
5	CasMVSNet(…		28.75	17.73	19.20	CasMV…		
	Xiaodong Gu et al."Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching".arXiv 2019.							
6	UCS-Net (ba…		28.30	17.77	19.02	UCS-Net		
	Shuo Cheng et al."Deep stereo using adaptive thin volume representation with uncertainty awareness".CVPR 2020.							
7	算法cj		23.446456	17.03879	18.58486			
8	COLMAP (ba…		**40.57**	13.14	17.90	COLMAP		
	Johannes Lutz Schönberger and Jan-Michael Frahm."Structure-from-Motion Revisited".CVPR 2016.							

Our method rank 2nd on Track Reconstruction

GIGAVISION

Tricks

- **Precision & Recall Balance +1.0**

$$F(d) = \frac{2P(d)R(d)}{P(d) + R(d)}.$$

In order to get a high F-score, Precision or Recall should not be too low.

- **Remove useless views +0.1**

- **Point cloud filtering +0.5**

We use Backbone point clouds sampled on the buildings to automatically crop the final results.

- **Point cloud complement +0.2**

We merge Colmap points with high precision and low recall to complete our results.

GIGAVISION

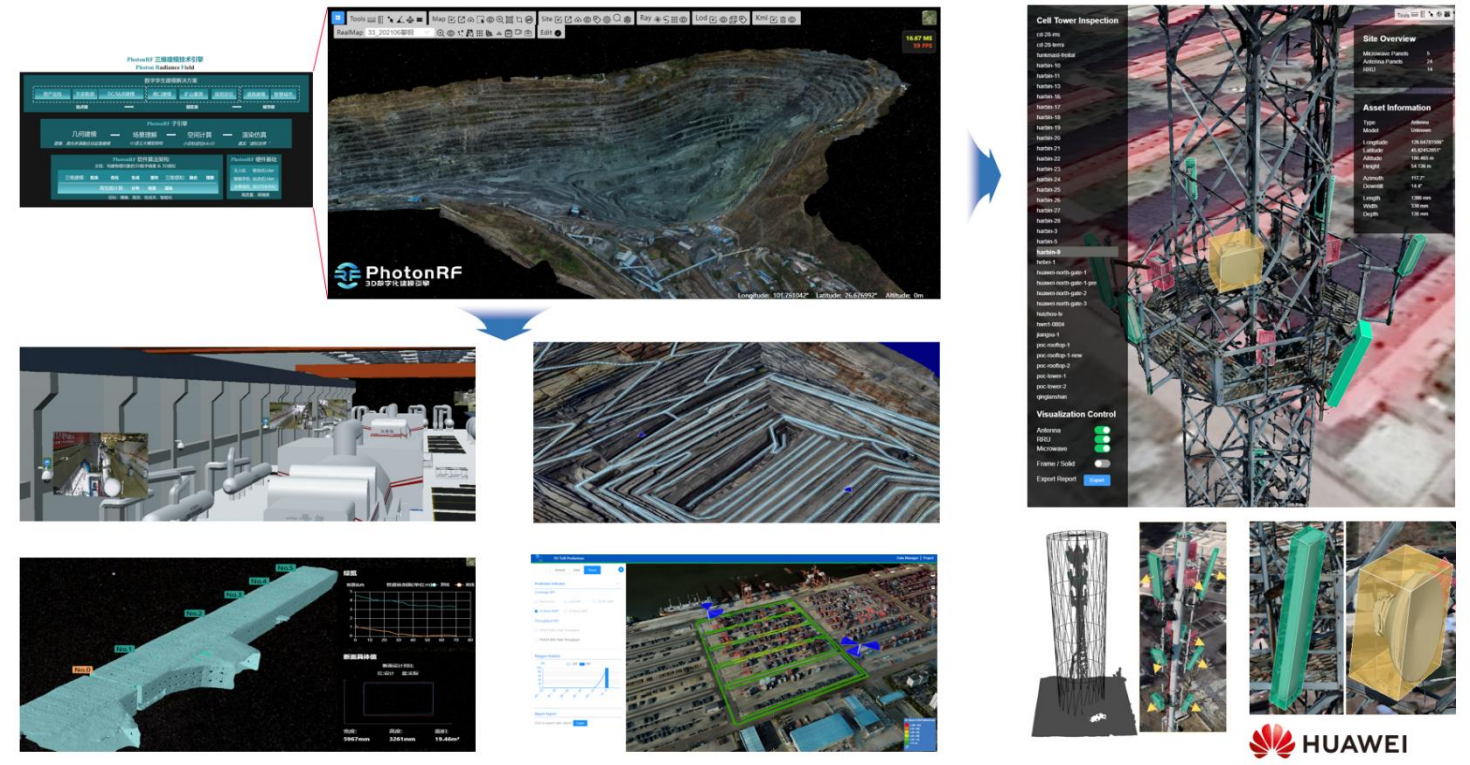
Conclusion

- Our method took the **second place** using a multi-scale patchmatch based framework, in which **multi-scale** feature matching, **view selection** and dynamic/semantic **fusion algorithms** play a key role.
- We‘ve tried some deep learning methods, but they don’t work well and can‘t produce high-resolution depth maps. However, theoretically, **learning based MVS methods** are expected to achieve better results.

GIGAVISION

Invitation

Digital Twin Lab, Huawei



Our team focus on cutting-edge technology research and engine development of **image/LiDAR 3D reconstruction** and **2/3D semantic understanding** for solving technical problems such as **environment 3D modeling** and perception in **5G network simulation**.

Contact: Hong.Shen233@huawei.com

Welcome to join us!

GIGAVISION

THANKS !