3D Reconstruction in Challenging Sparse View Setup

Nischal Maharjan
Universität des Saarlandes
ETH Student Summer Research Fellowship 2025

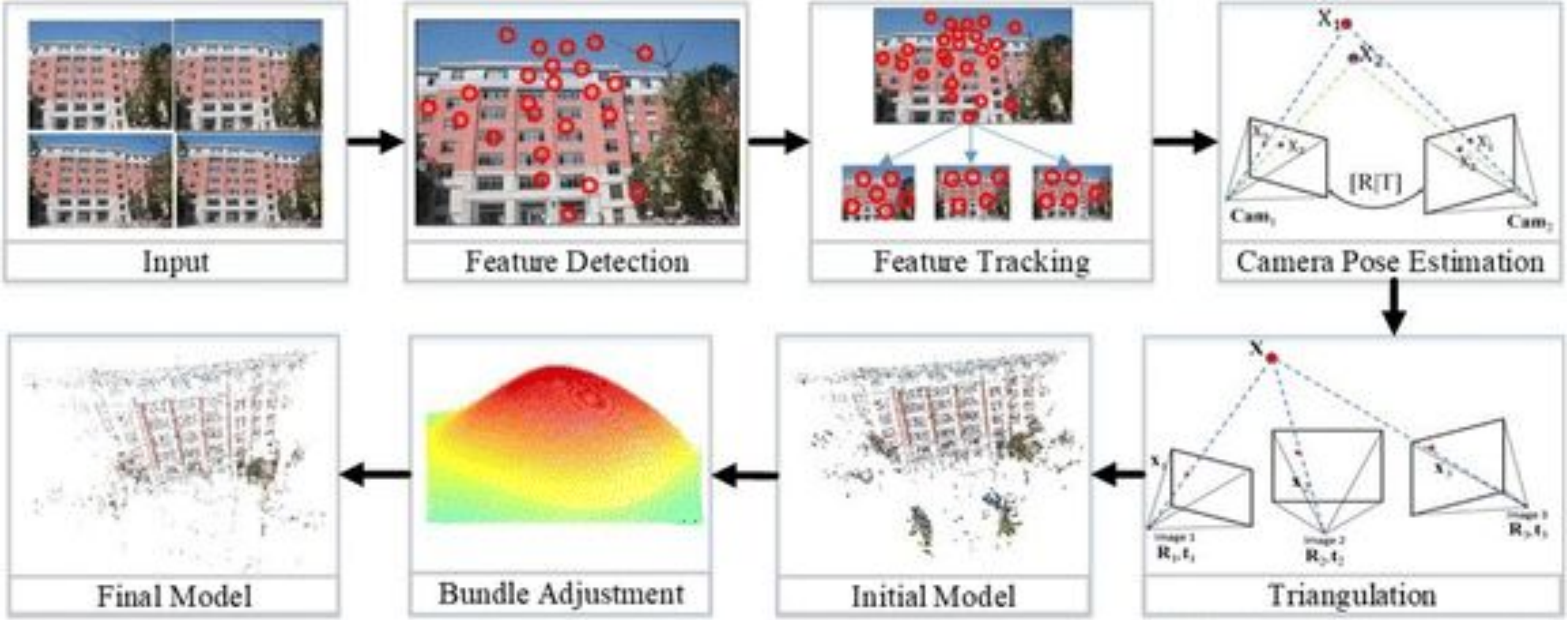
Supervisor: **Sergey Prokudin and Yutong Chen**
Computer Vision and Learning Lab

Introduction

3d
Reconstruction

Structure from Motion

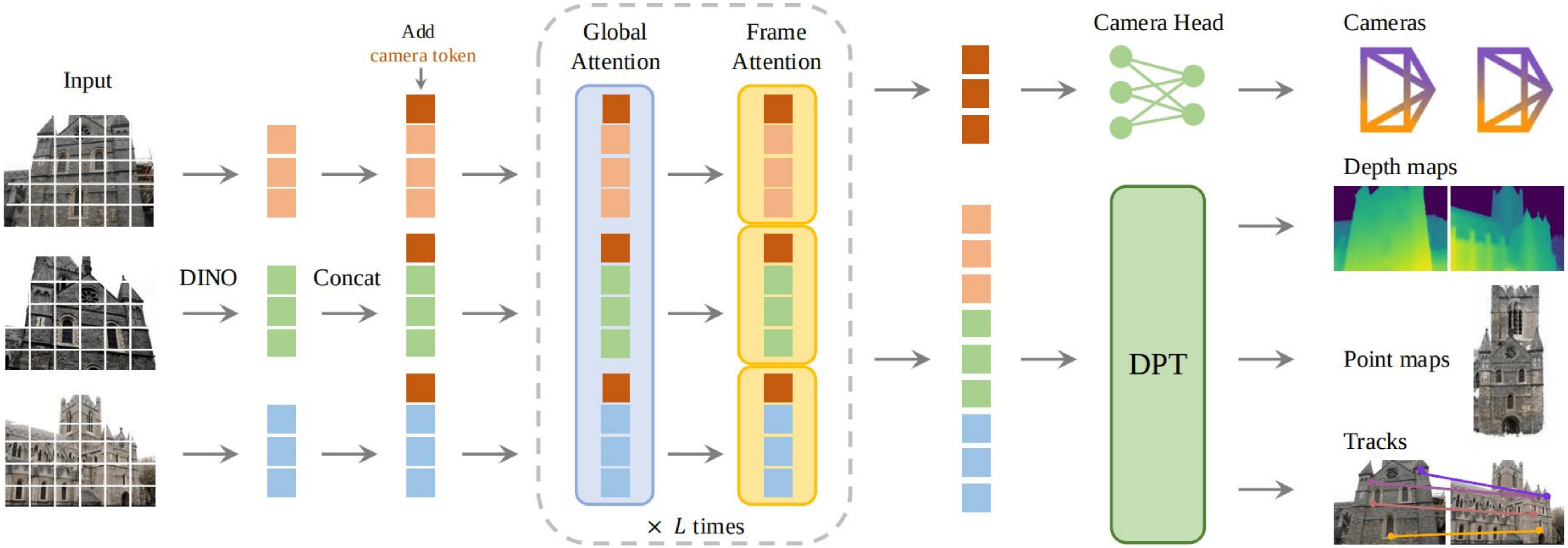
Structure from Motion Pipeline



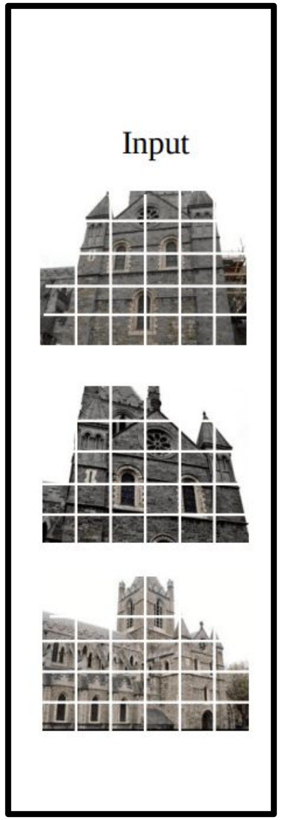
Limitations

- extreme viewpoint changes in low-overlap,
- low-parallax or high-symmetry scenarios.
- Scene without texture makes it difficult to detect feature points

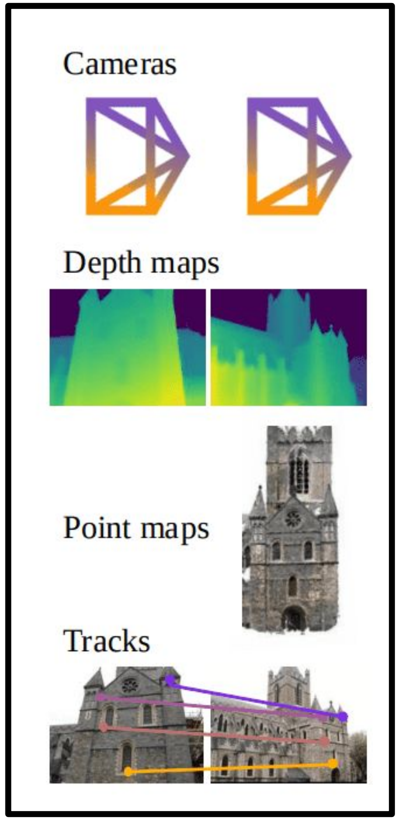
VGGT Model- Predicts camera parameters and point maps



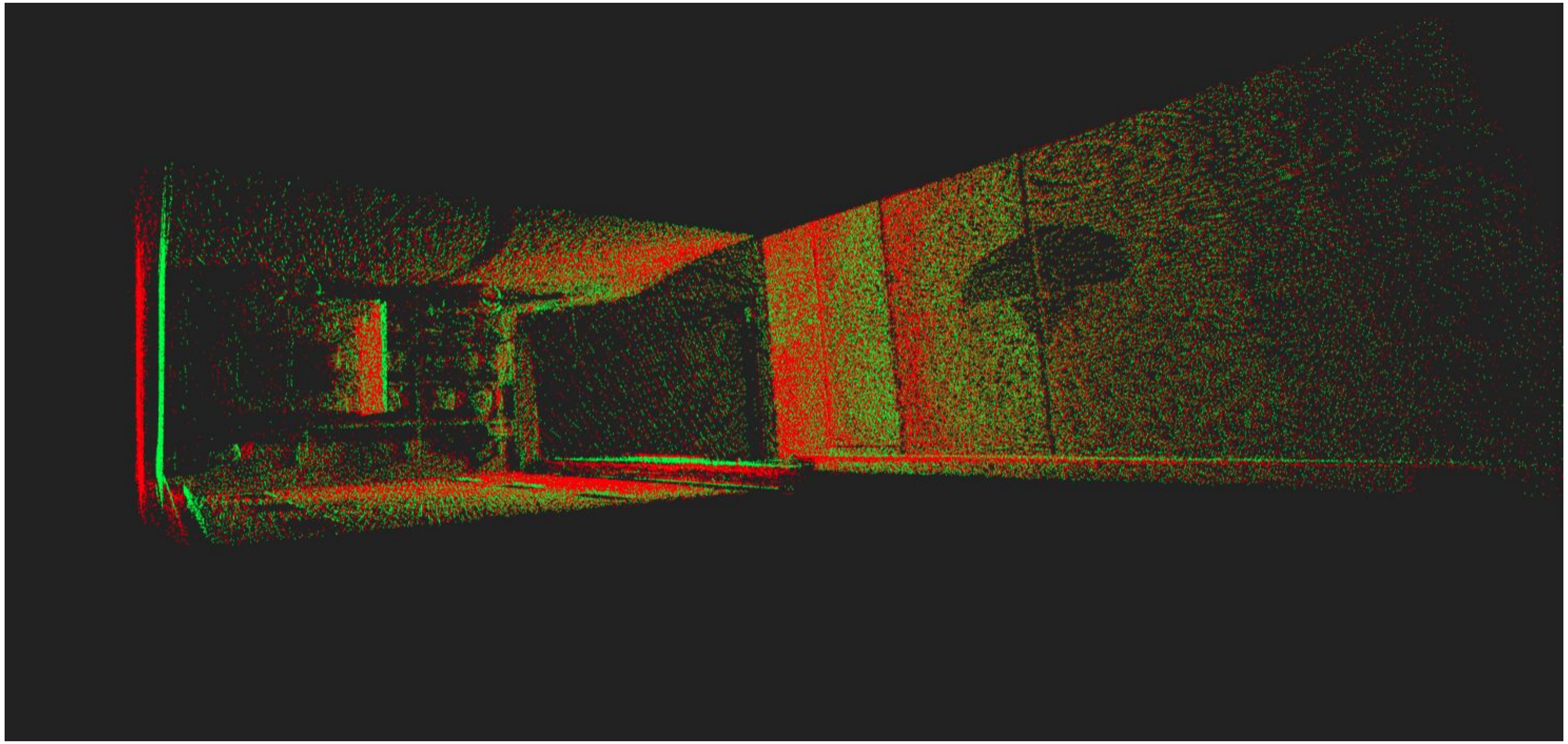
VGGT Model- Predicts camera parameters and point maps



- Given Input Images
- It outputs
  - Camera Parameters
  - Depth Maps
  - Point Maps
  - Tracking points



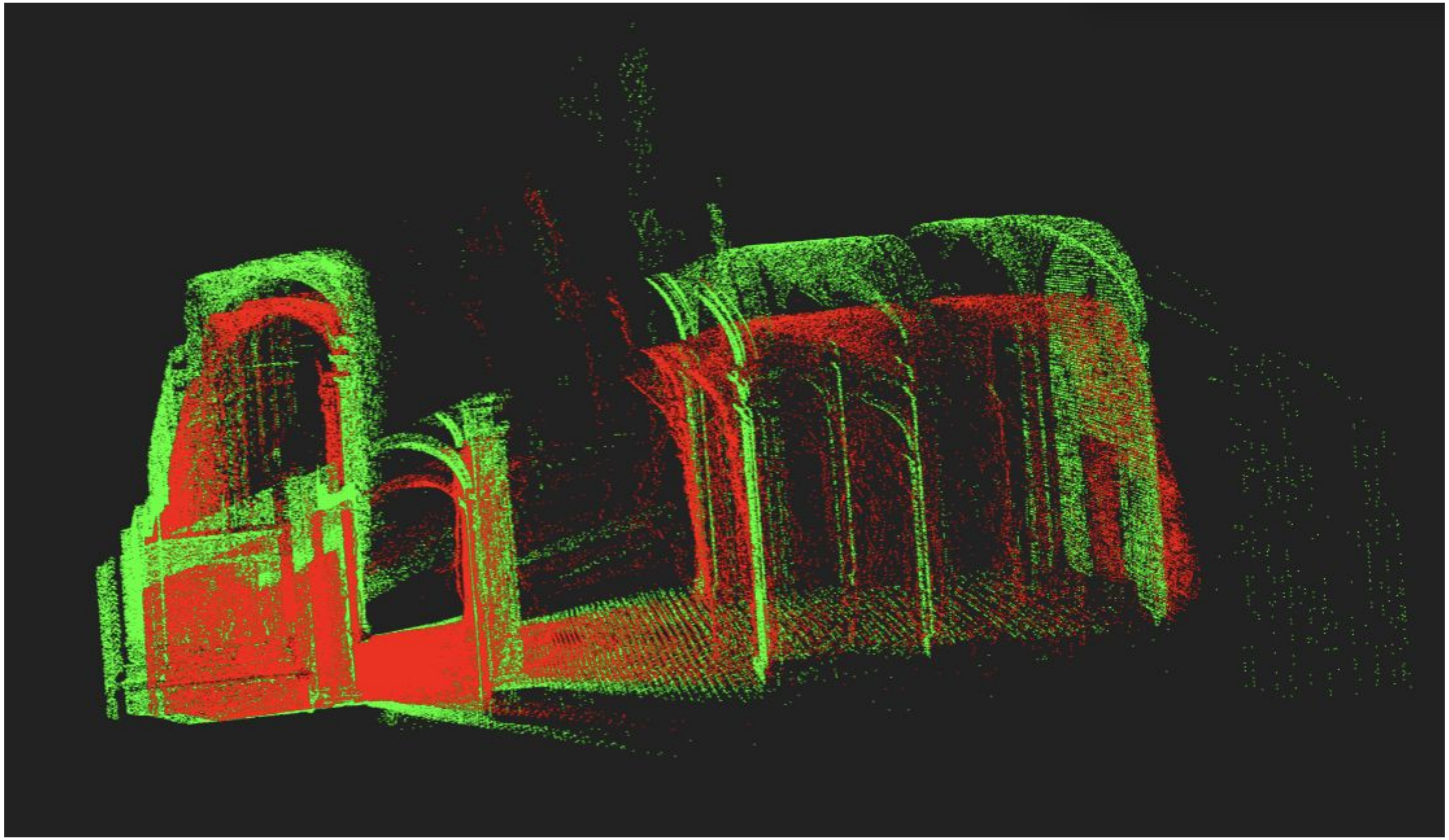
Drawbacks of VGGT



Green - Ground truth
Red- Prediction

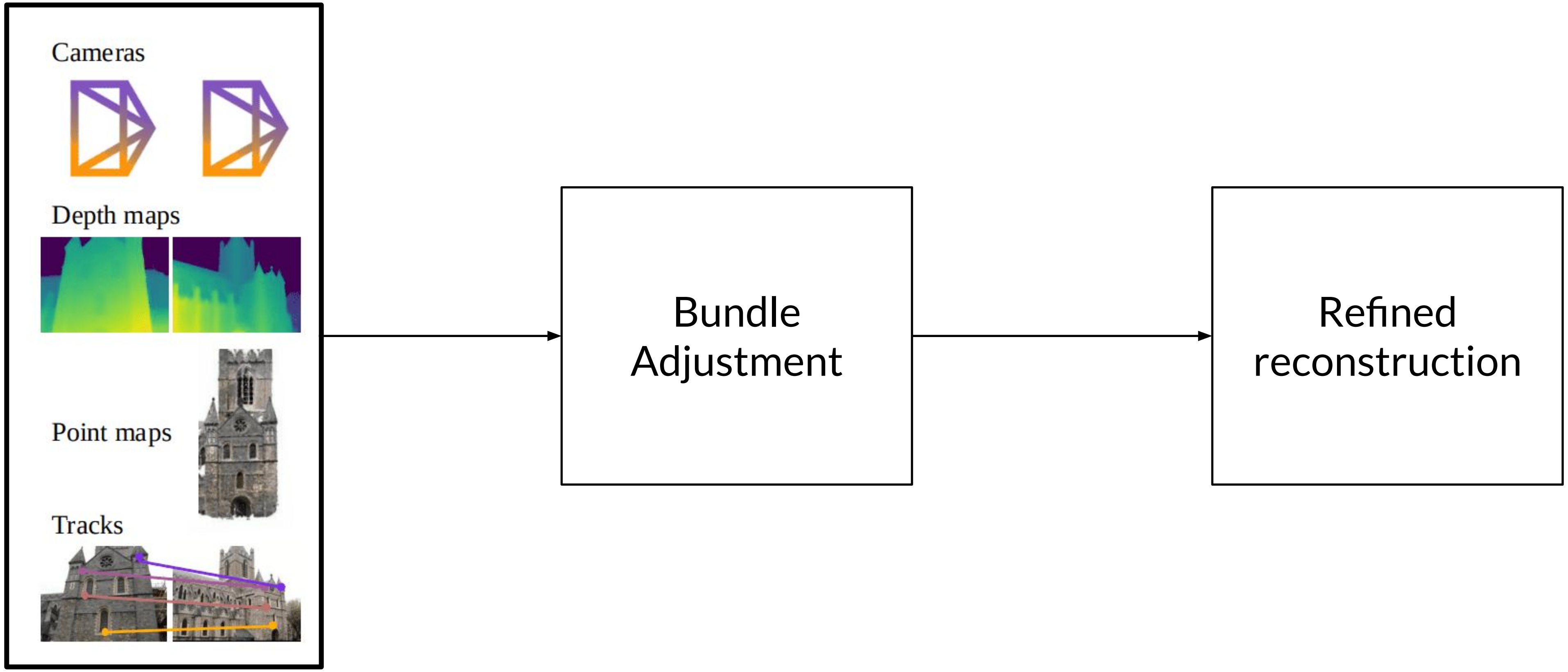
The Structure is correct but lacks global alignment

Drawbacks of VGGT



Green - Ground truth
Red- Prediction

Goal: VGGT+BA (Use VGGT predictions as prior for BA)



Bundle Adjustment

- Bundle Adjustment minimizes the **reprojection error**

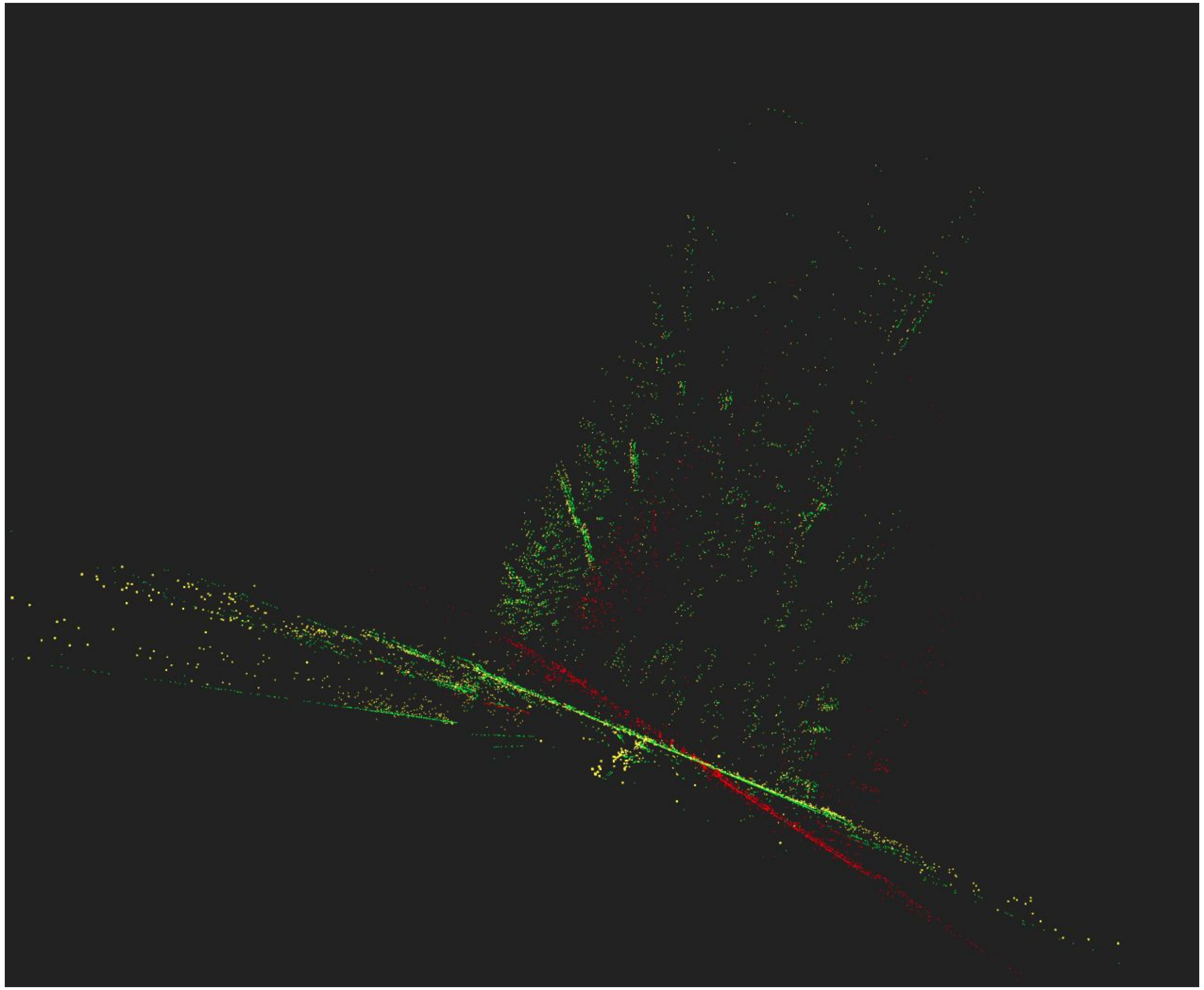
$$reproj\_error = x_j^i - P_i X_j$$

where x_j^i is the 2d point corresponding to X_j in i^{th} viewpoint

and $P_i = K_i[R_i|t_i]$ is the projection matrix of i^{th} camera view

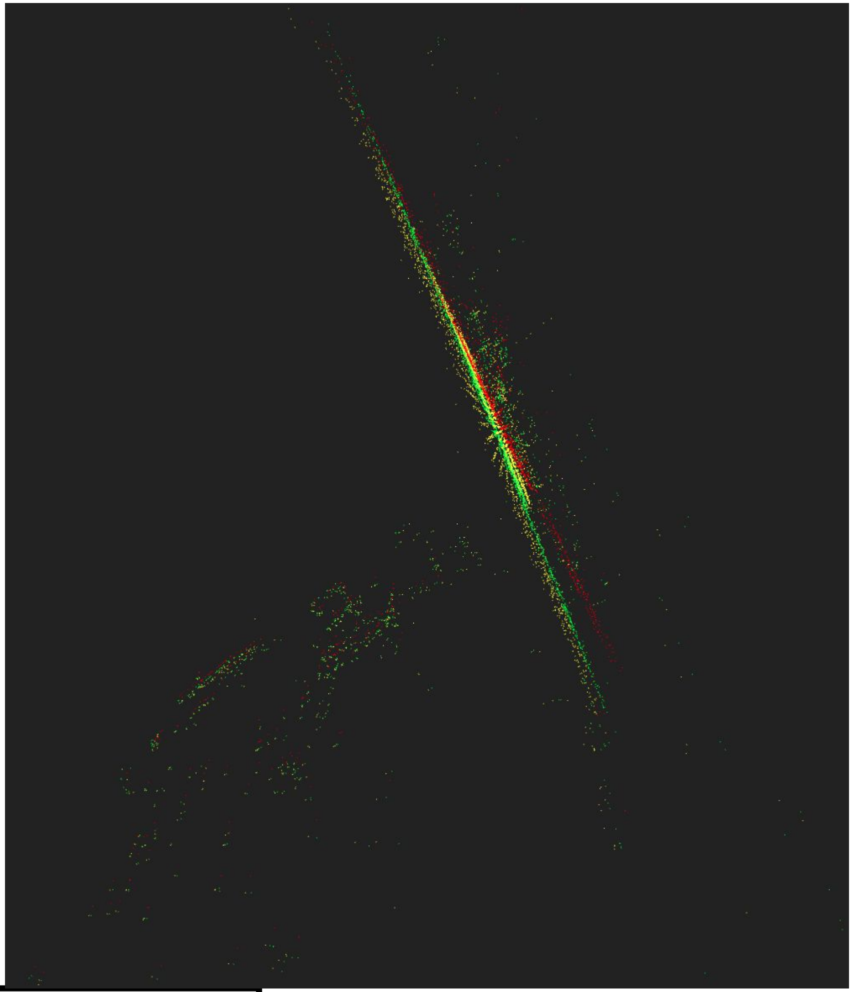
$$\boxed{P_i, X_j = argmin_{P_i, X_j} \sum_{i,j} ||x_j^i - P_i X_j||^2}$$

VGGT+BA



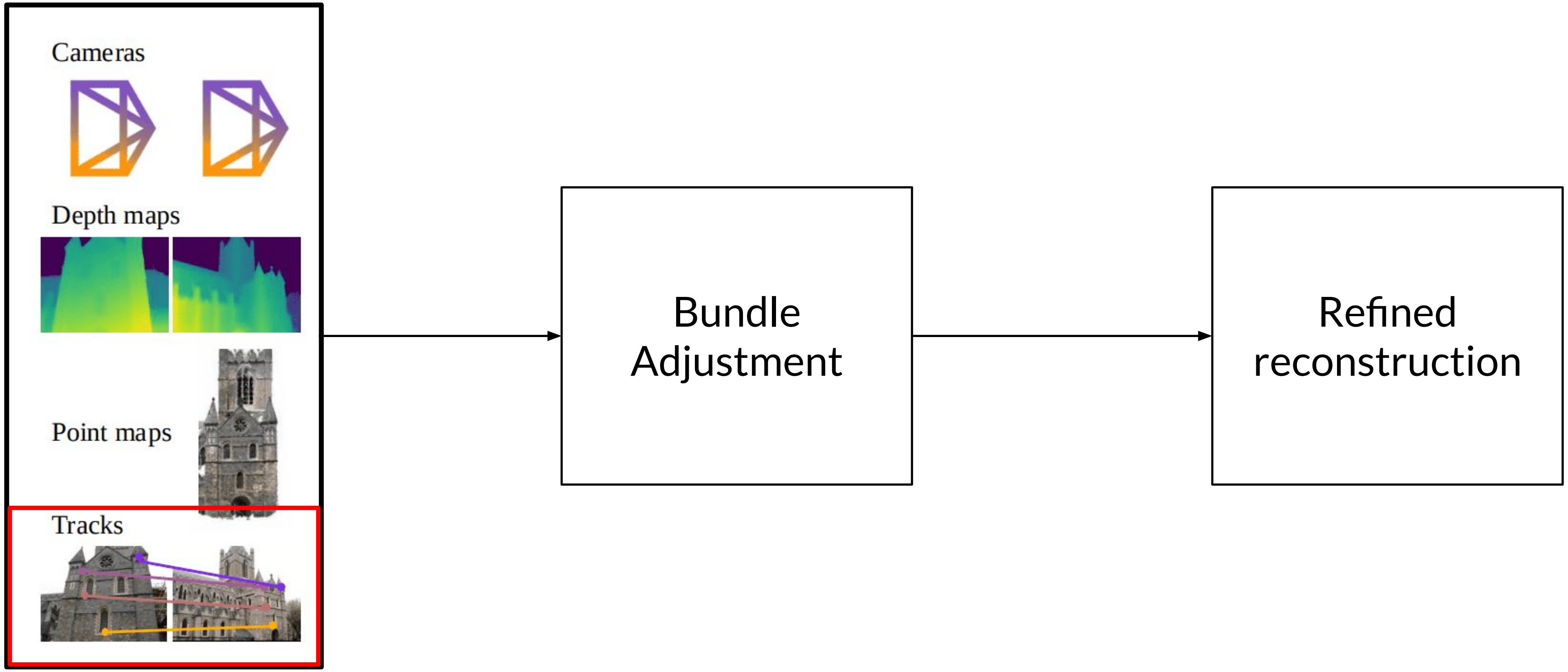
Green - Ground truth
Red- VGGT Prediction
Yellow - VGGT + BA

VGGT+BA

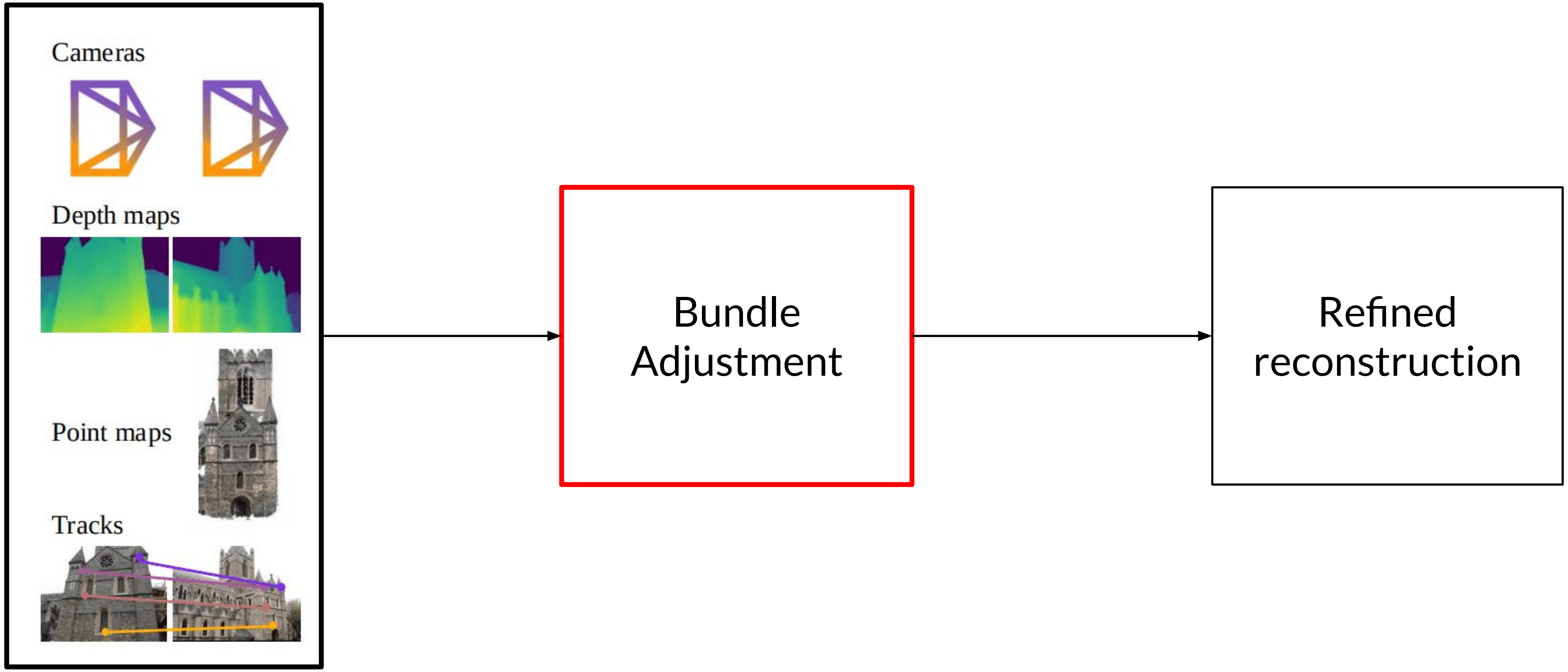


Green - Ground truth
Red- VGGT Prediction
Yellow - VGGT + BA

Improve the inputs to the BA: Input tracks



Improve the BA block



Experiments

1. **Inputs**
   a. VGGSfM vs MASt3R tracking module
   b. Effect of query points
   c. Filtering Correspondences

2. **BA parameters**
   a. Reapplying BA
   b. Loss Fuction

Metrics

1. Camera Metric
    a. Intrinsics -> error in field of view
    b. Extrinsic -> Computes how accurately the rotation and translation are estimated

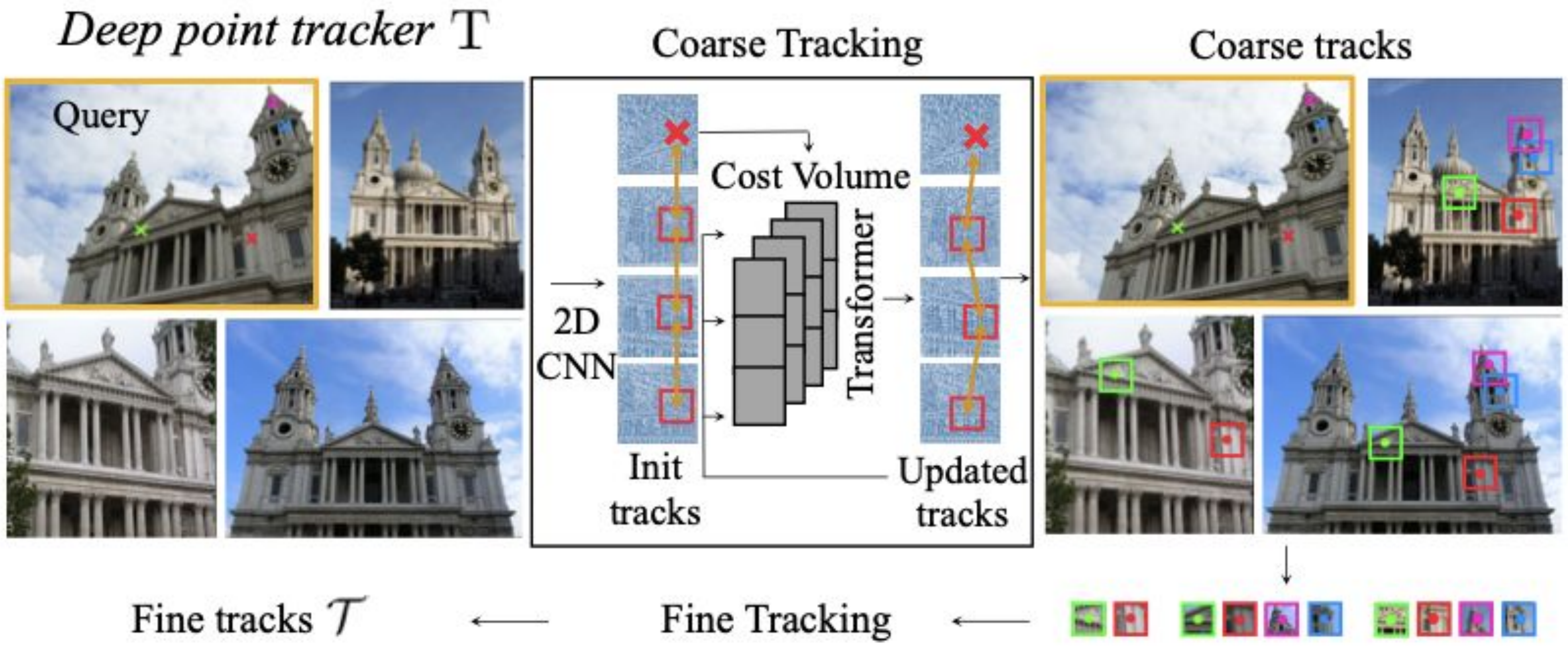
2. 3D metric
    a. Error in position of point clouds
    b. Accuracy of points

3. Tracking metric
    a. Tracking error
    b. Tracking statistics

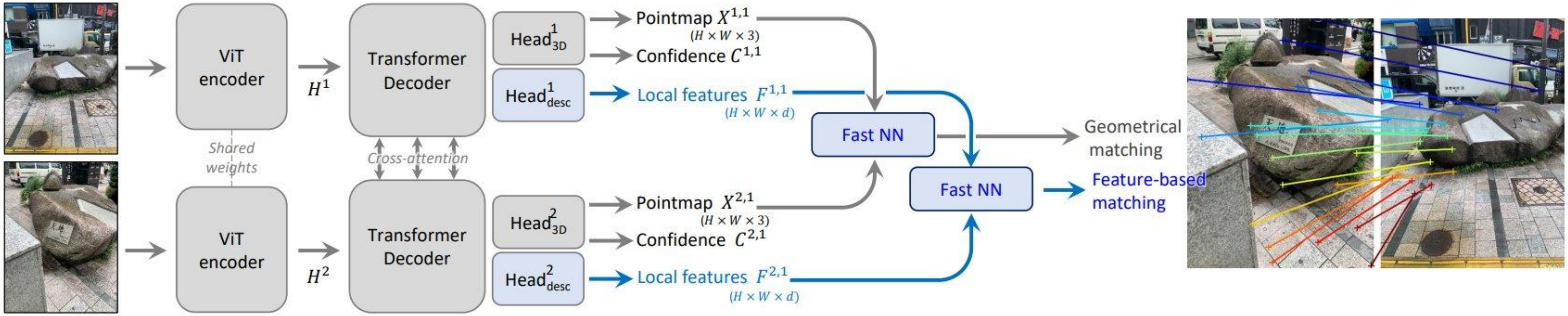
ETH3D Dataset

VGGSfM Tracks

- Gets embeddings features for query points using 2d CNN
- Creates Cost volume Pyramid
- Transformer to update tracks
- Coarse to fine tracking



MASt3R Tracks



- Build upon DUSt3R architecture
- Specifically targeted to finding dense matches
- Limitation: Used for Pair-wise match estimation

VGGSfM Tracks Vs MASt3R tracks

Camera Metrics (Support = 109)		VGGSfM	MASt3R
Extrinsics	auc@01(%) ↑	**74.90**	71.13
	auc@03(%) ↑	**83.38**	80.23
	auc@05(%) ↑	**86.78**	84.26
	auc@10(%) ↑	**90.49**	88.96
	auc@20(%) ↑	**93.42**	92.36
	auc@30(%) ↑	**94.77**	94.03
Intrinsics	fovx error(deg) ↓	0.98	**0.96**
	fovy error(deg) ↓	1.60	**1.00**

VGGSfM tracks lead to better extrinsics metrics whereas MASt3R tracks has better intrinsic metrics

VGGSfM Tracks Vs MASt3R tracks

3D Metrics (Support = 109)		VGGSfM	MASt3R
Error	rmse_mean(cm) ↓	899.36	**434.32**
	rmse_median(cm) ↓	**6.69**	10.56
AUC	auc@02cm(%) ↑	**20.67**	16.59
	auc@04cm(%) ↑	**32.28**	27.97
	auc@06cm(%) ↑	**40.23**	35.97
	auc@08cm(%) ↑	**46.20**	42.10
	auc@10cm(%) ↑	**50.92**	47.00

VGGSfM tracks are better than MASt3R tracks

VGGSfM Tracks Vs MASt3R tracks

Tracking Metrics (Support = 109)		VGGSfM	MASt3R
Track error	tracking_error/mean ↓	**2.13**	4.07
	tracking_error/median ↓	**0.90**	2.14
Track statistics	mean_track_length ↑	3.79	3.85
	median_track_length ↑	3.89	3.87
	max_track_length ↑	7.07	7.11
	full_track_percentage ↑	6.68	5.43

VGGSfM tracks are better than MASt3R tracks

Effect of query points

Camera Metrics (Support = 109)		max_query_pts = 2048	max_query pts = inf
Extrinsics	auc@01(%) ↑	70.36	70.27
	auc@03(%) ↑	79.46	79.34
	auc@05(%) ↑	83.47	83.14
	auc@10(%) ↑	88.14	87.40
	auc@20(%) ↑	91.68	91.02
	auc@30(%) ↑	93.53	92.83
Intrinsics	fovx error(deg) ↓	1.08	1.12
	fovy error(deg) ↓	1.12	1.19

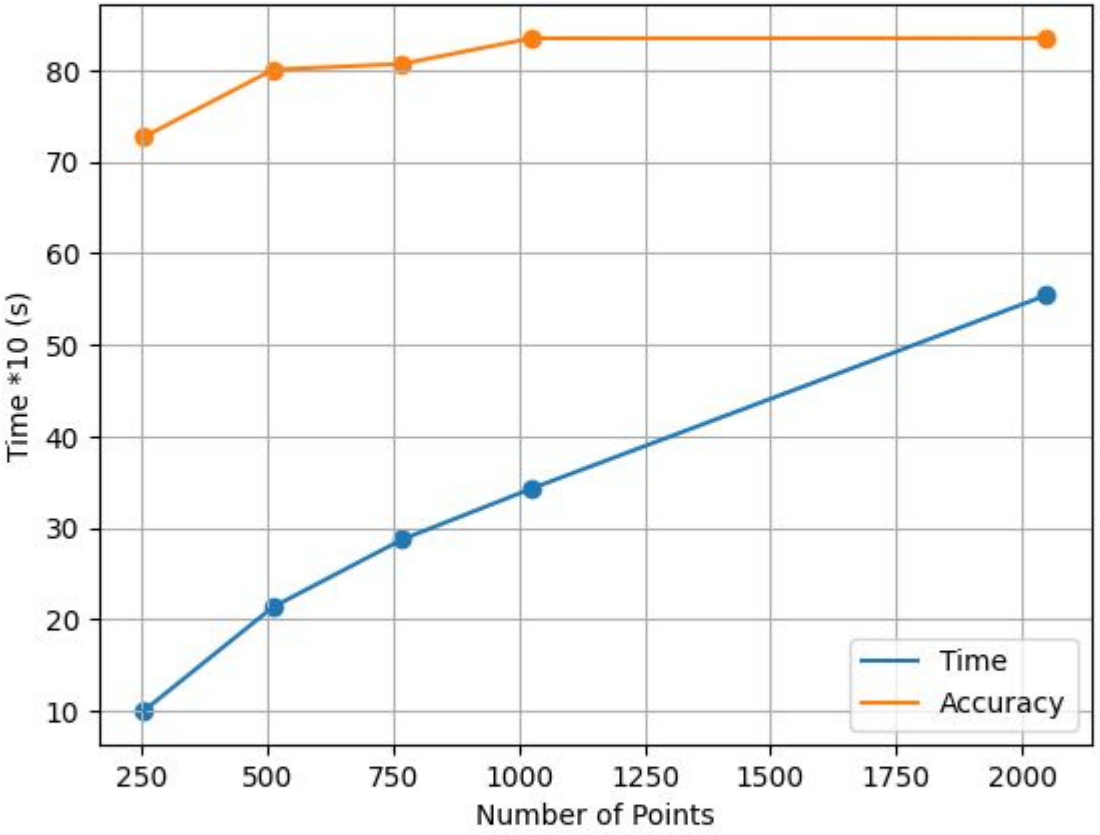
Change in metrics were very insignificant but change in time complexity was significant

Effect of query points

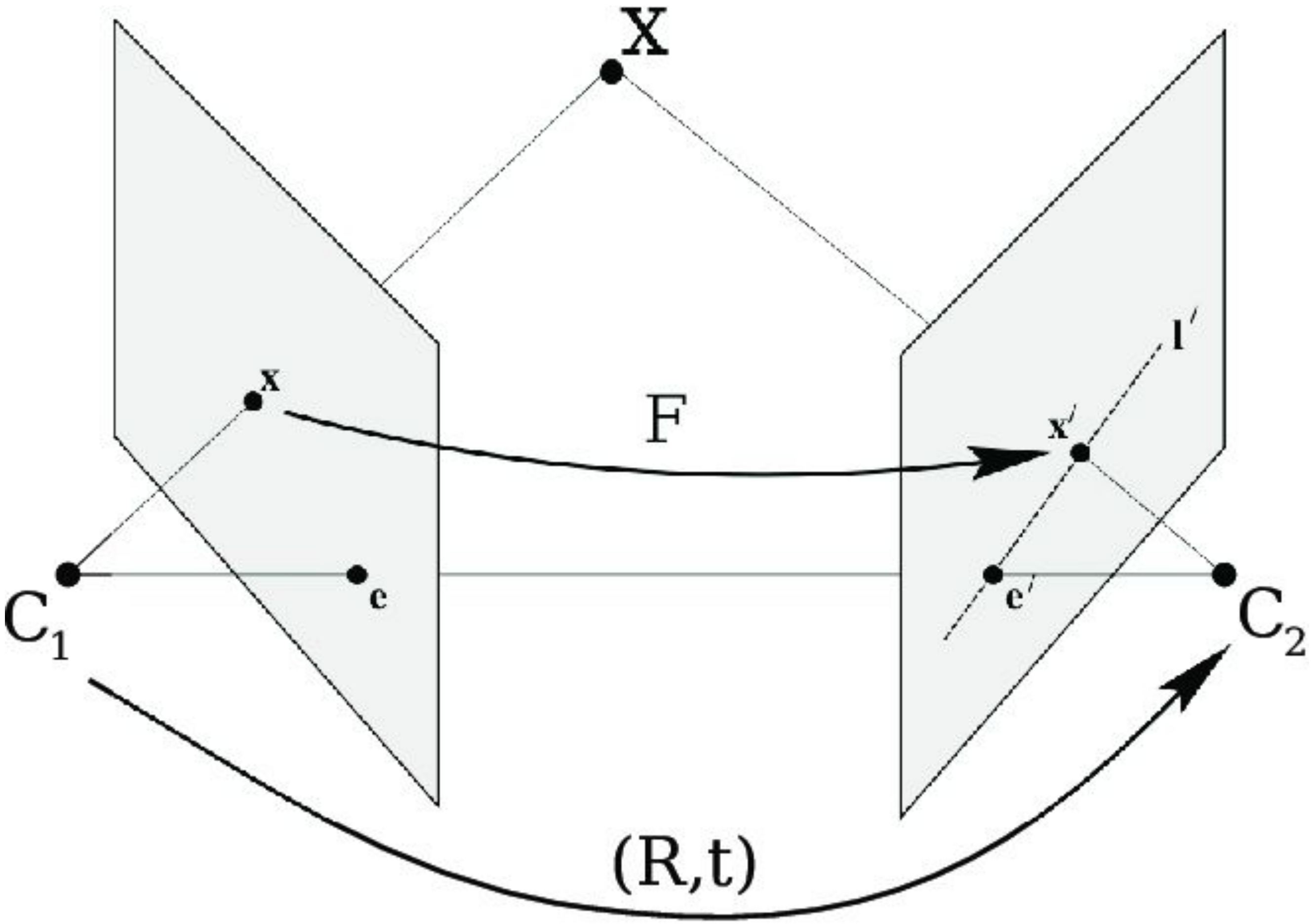
- Investigated the camera metric accuracy and time complexity for BA over number of query pts

- Accuracy has slight decrease but time taken for BA has significant drop when less points are used

- In case **robust triangulation method** exists estimating camera metric with less points decreases time without significant drop in performance



Epipolar Constraint



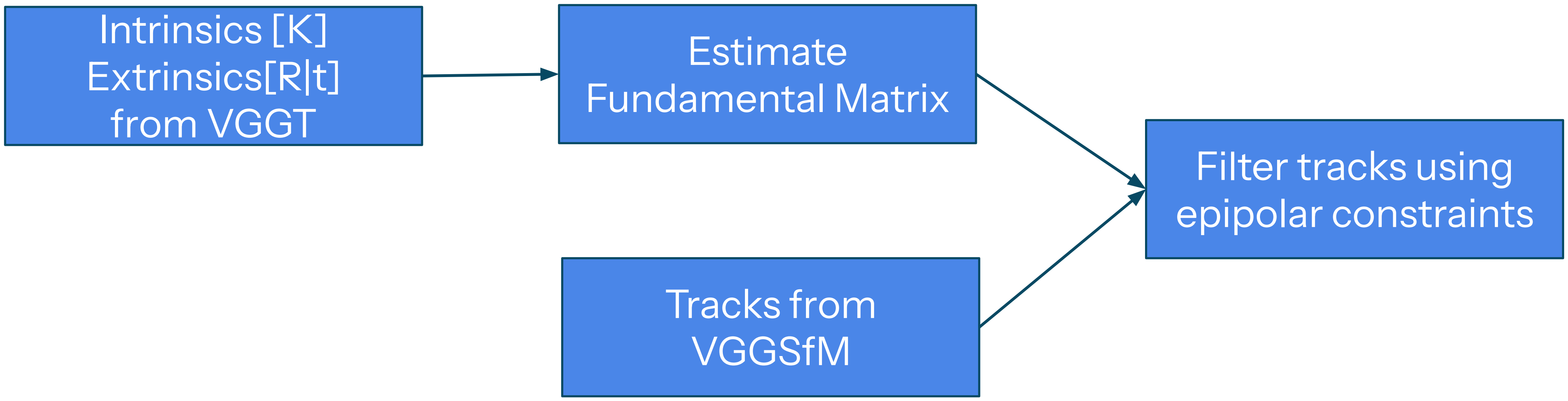
$$\boxed{x_2^T F x_1 = 0}$$

$$F = K^{-T} E K^{-1}$$

$$E = [t]_\times R$$

Fundamental matrix can be estimated from intrinsics and extrinsic parameters

Filtering Correspondences



Filtering Correspondences

Tracking Metrics (Support = 94) (Cauchy loss)		without filter	with filter
Track error	tracking_error/mean ↓	2.11	**1.70**
	tracking_error/median ↓	0.89	**0.82**
Track statistics	mean_track_length ↑	**3.99**	2.77
	median_track_length ↑	**4.12**	2.68
	max_track_length ↑	**7.29**	6.52
	full_track_percentage ↑	**7.95**	3.12

Tracking error decreased however the track length also was decreased

Filtering Correspondences

Camera Metrics (Support = 94) (Cauchy loss)		without filter	with filter
Extrinsics	auc@01(%) ↑	**87.95**	75.07
	auc@03(%) ↑	**93.445**	83.54
	auc@05(%) ↑	**95.31**	87.16
	auc@10(%) ↑	**97.17**	90.95
	auc@20(%) ↑	**98.39**	93.90
	auc@30(%) ↑	**98.88**	95.26
Intrinsics	fovx error(deg) ↓	**0.49**	0.84
	fovy error(deg) ↓	**1.05**	1.25

Using filter didn't improve the performance even though the tracking error was improved

Filtering Correspondences

3D Metrics (Support = 94) (Cauchy loss)		without filter	with filter
Error	rmse_mean(cm) ↓	4444.10	**2313.32**
	rmse_median(cm) ↓	**5.07**	17.44
AUC	auc@02cm(%) ↑	**24.03**	22.74
	auc@04cm(%) ↑	**36.06**	34.99
	auc@06cm(%) ↑	**44.14**	43.36
	auc@08cm(%) ↑	**50.20**	49.60
	auc@10cm(%) ↑	**54.97**	54.48

Using filter didn't improve the performance even though the tracking error was improved

Re-applying BA (ReBA)

- Filter 3D points based upon reprojection error and triangulation angle
- Re-apply BA on filtered set of points

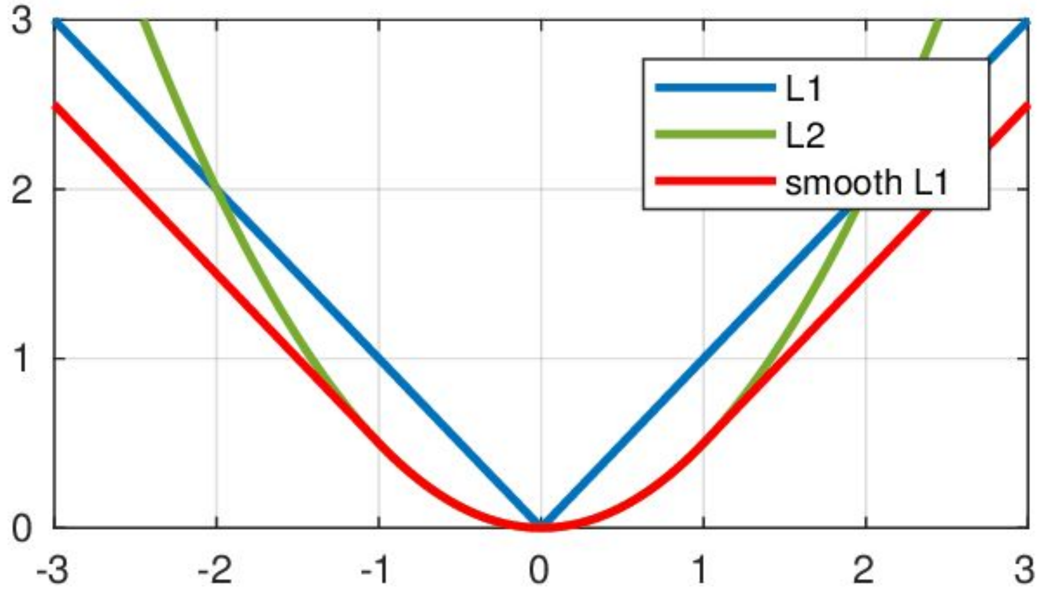
Re-applying BA (ReBA)

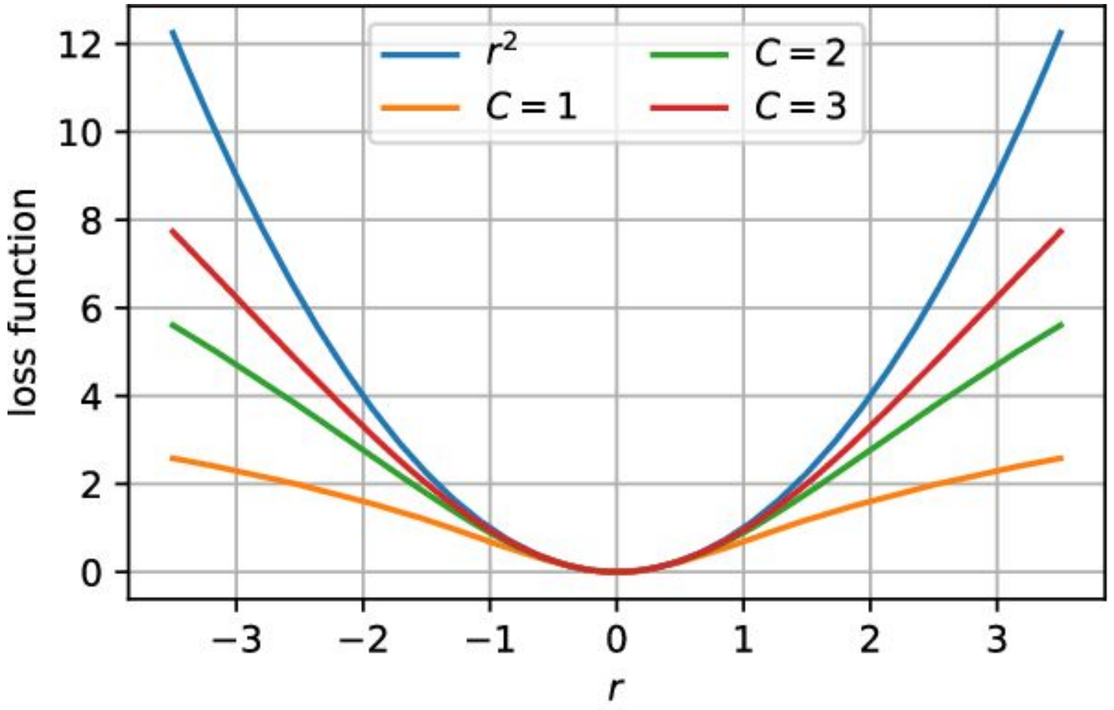
3D Metrics (Support = 109)		First BA	ReBA
Error	rmse_mean(cm) ↓	**13.65**	15.29
	rmse_median(cm) ↓	**6.59**	6.62
AUC	auc@02cm(%) ↑	22.95	**23.27**
	auc@04cm(%) ↑	34.98	**35.14**
	auc@06cm(%) ↑	43.08	**43.16**
	auc@08cm(%) ↑	49.11	**49.12**
	auc@10cm(%) ↑	**53.82**	53.80

Reapplying BA there is slight improvement in accuracy but not significant

Loss Functions

- Trivial (L2 loss)
- Soft_L1 loss
- Robust (Cauchy loss)





Loss Functions

Camera Metrics (Support = 116)		L2 loss	Soft_L1 loss	Cauchy loss
Extrinsics	auc@01(%) ↑	72.44	76.63	**78.60**
	auc@03(%) ↑	81.36	84.09	**85.44**
	auc@05(%) ↑	85.04	86.98	**88.10**
	auc@10(%) ↑	89.11	90.27	**90.95**
	auc@20(%) ↑	92.31	93.04	**93.40**
	auc@30(%) ↑	93.83	94.42	**94.67**
Intrinsics	fovx error(deg) ↓	1.16	1.07	**0.99**
	fovy error(deg) ↓	1.64	1.41	**1.22**

Cauchy loss performed better than other

Loss Functions

3D Metrics (Support = 116)		L2 Loss	Soft_L1 Loss	Cauchy Loss
Error	rmse_mean(cm) ↓	**847.48**	2121.21	3621.12
	rmse_median(cm) ↓	7.55	8.87	**6.31**
AUC	auc@02cm(%) ↑	20.12	21.80	**22.96**
	auc@04cm(%) ↑	31.68	33.40	**34.74**
	auc@06cm(%) ↑	39.56	41.30	**42.63**
	auc@08cm(%) ↑	45.48	47.24	**48.53**
	auc@10cm(%) ↑	50.16	51.92	**53.19**

Cauchy loss performed better than other

Effect of different Loss Scales

Camera Metrics (Support = 116)		Cauchy loss scales						
		0.05	**0.1**	**0.2**	**0.5**	**1**	**2**	**3**
Extrinsics	auc@01(%) ↑	**84.54**	84.42	83.65	81.40	78.60	76.44	75.36
	auc@03(%) ↑	**88.98**	88.42	88.39	87.34	85.44	84.01	83.29
	auc@05(%) ↑	**90.76**	90.23	90.28	89.54	88.10	86.93	86.45
	auc@10(%) ↑	**92.92**	92.43	92.47	92.01	90.95	90.24	89.99
	auc@20(%) ↑	**94.80**	94.35	94.33	94.06	93.40	92.98	92.89
	auc@30(%) ↑	**95.87**	95.39	95.35	95.13	94.67	94.39	94.32
Intrinsics	fovx error(deg) ↓	**0.66**	0.66	0.69	0.85	0.99	1.07	1.10
	fovy error(deg) ↓	**0.81**	0.81	0.93	1.12	1.22	1.40	1.46

Performance increases monotonously as scale decreases

Effect of different Loss Scales

Camera Metrics (Support = 116)		**Cauchy loss scales**						
		0.05	**0.1**	**0.2**	**0.5**	**1**	**2**	**3**
Error	rmse_median(cm) ↓	**5.49**	5.65	5.59	5.82	6.31	7.05	9.11
AUC	auc@02cm(%) ↑	24.93	24.93	**25.07**	24.24	22.96	22.04	21.38
	auc@04cm(%) ↑	**37.13**	37.13	37.03	36.06	34.74	33.70	32.93
	auc@06cm(%) ↑	**45.17**	45.16	45.00	43.98	42.63	41.61	40.79
	auc@08cm(%) ↑	**51.08**	51.06	50.92	49.87	48.53	47.53	46.72
	auc@10cm(%) ↑	**55.68**	55.64	55.53	54.48	53.19	52.18	51.40

Performance increases as scale decreases

The trend is similar for Soft L1 loss as well

Further Enhancements

- Completing and merging tracks could improve the final results.

- Pixel-Perfect SfM in order to refine the keypoint for better tracks.

Conclusion

- Incorporating VGGT + BA helps in global alignment
- VGGSfM tracks has better reconstruction than MASt3R tracks
- A small tradeoff can be done for significant drop in time compleity by having small decrease in accuracy
- Track length seems to be important factor than the tracking accuracy.
- Re applying BA has slight improvement in performance
- Cauchy Loss has better performance than L2 and Soft L1 loss
- As scale decreases the results are better

THANK YOU!!