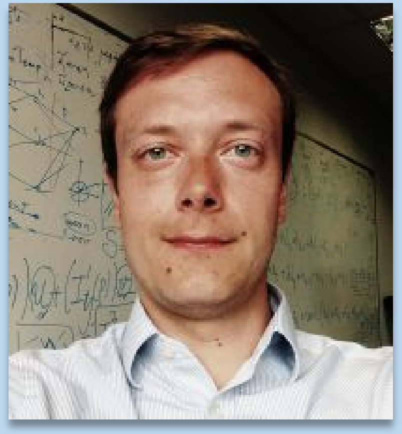
IMP: Iterative Matching and Pose Estimation with Adaptive Pooling

Fei Xue



Ignas Budvytis

Roberto Cipolla

UNIVERSITY OF CAMBRIDGE

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

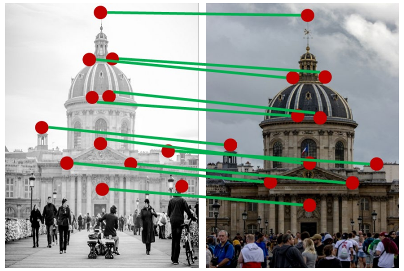
Preview of IMP

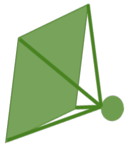
Input
Two sets of keypoints

Classic pipeline
Two separate steps

Output
Matches & Relative pose

Feature Matching → Pose estimation





Ignore the geometric connections
Slow
Inaccurate

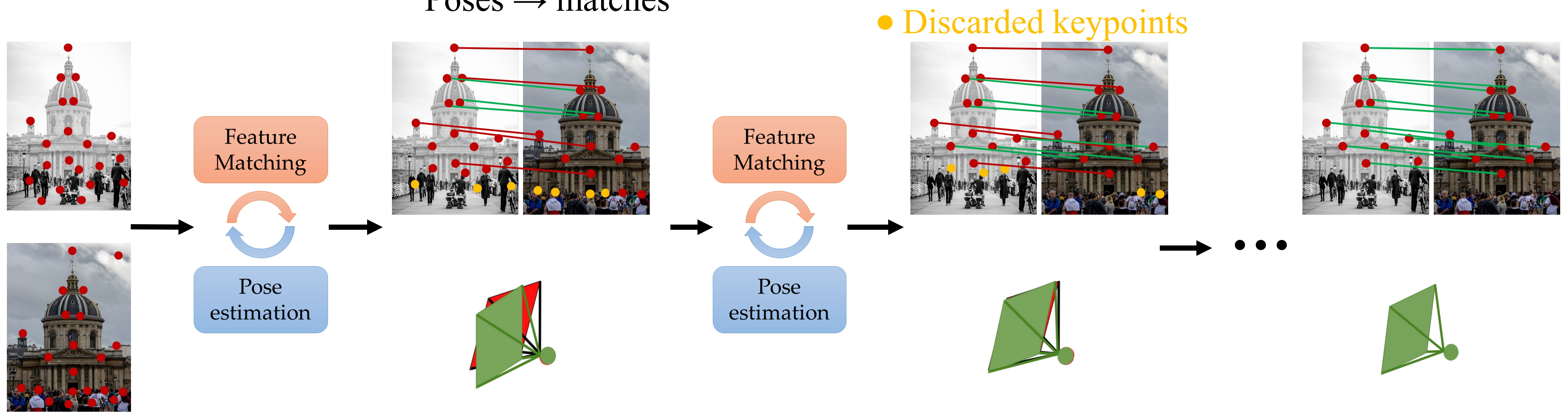
Preview of IMP

Input
Two sets of keypoints

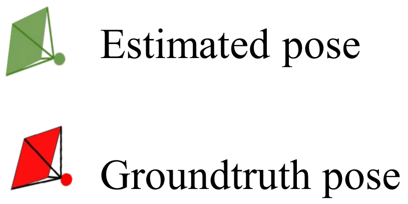
Iterative pipeline
Matches → poses
Poses → matches

Adaptive pooling
Discard useless keypoints

Output
Matches & Relative pose

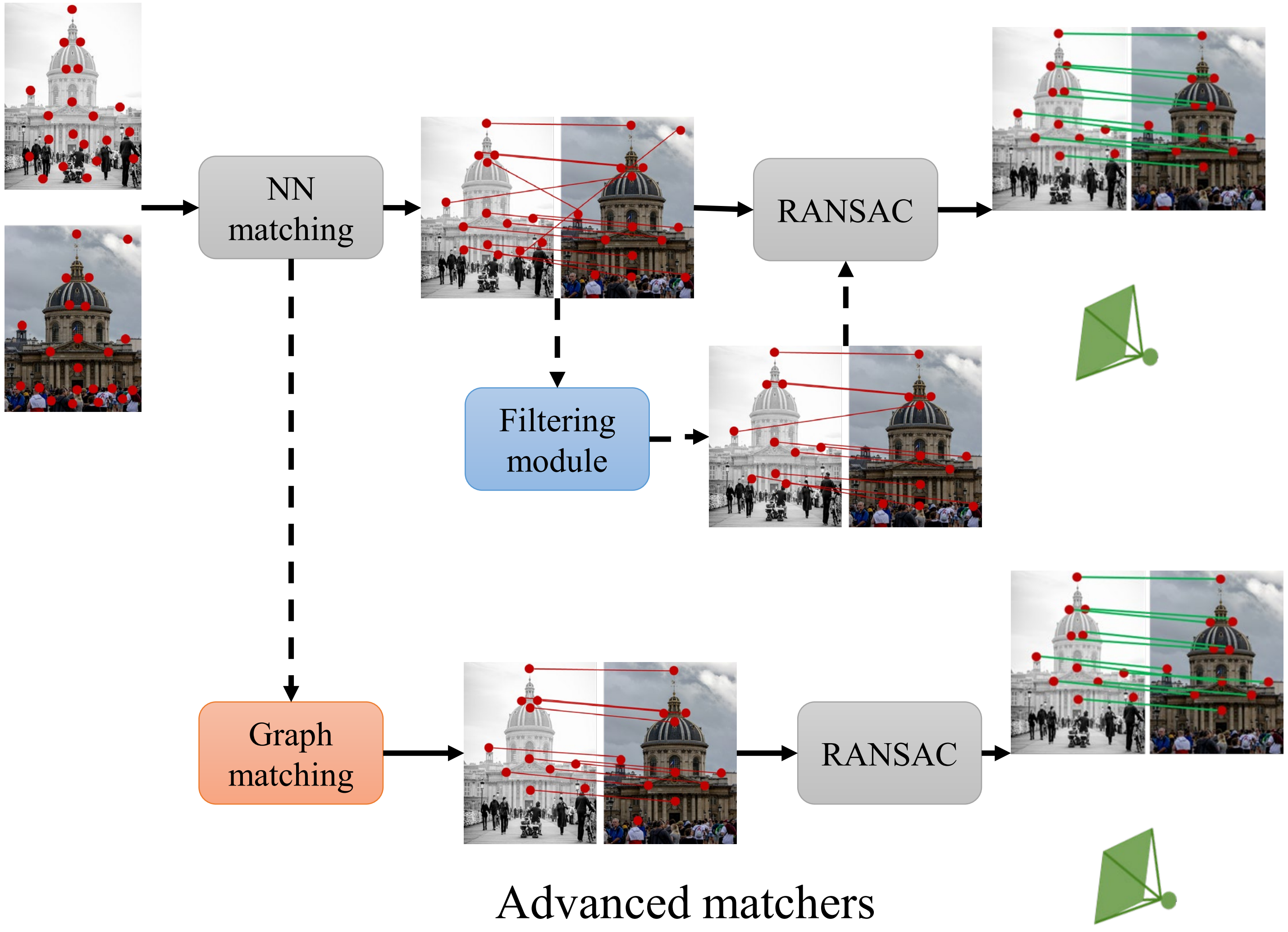


Retain the geometric connections
Faster
More accurate



Feature matching and pose estimation

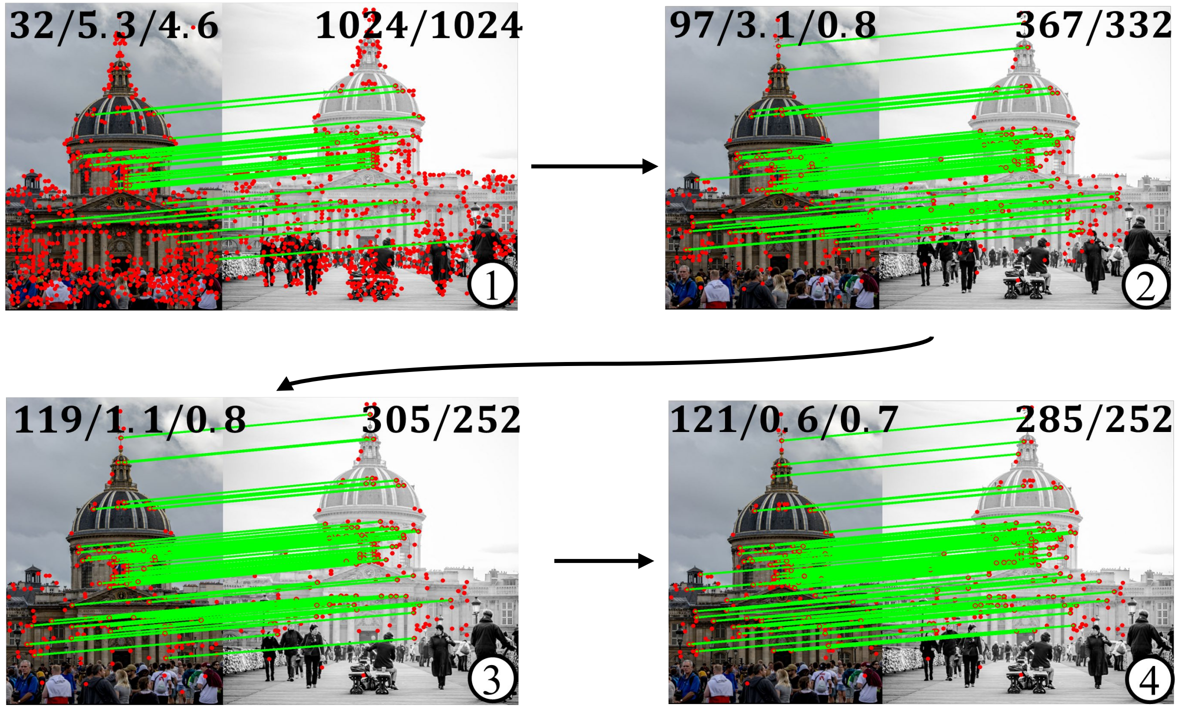
- **Traditional approaches**
  - Two separate steps
  - Slow & inaccurate
- **Outlier filtering**
  - Promising performance
  - Accuracy limited by initial matches
- **Advanced matchers**
  - Good accuracy
  - Quadratic time cost



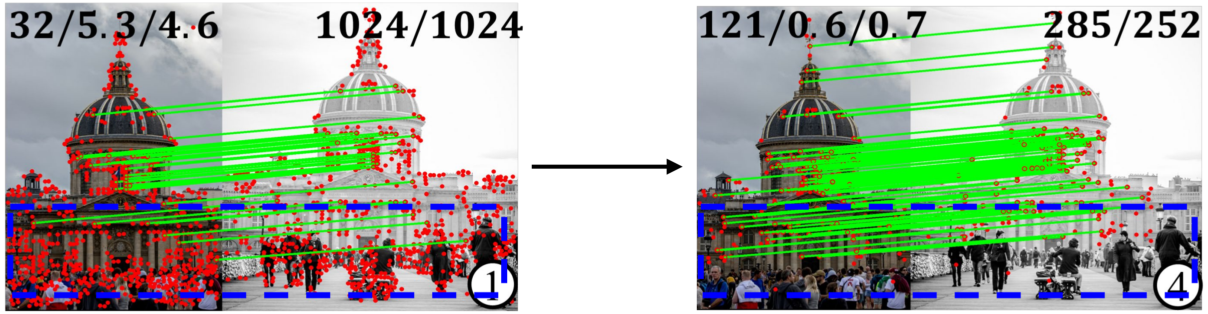
[1] Zhang et al., Learning two-view correspondences and geometry using order-aware network, ICCV 2019
[2] Sarlin et al., Superglue: Learning feature matching with graph neural networks, CVPR, 2020

Motivation

- **Geometric connections**
  - Several matches give a coarse pose
  - The pose guides the matching
- **Keypoints pooling**
  - Not all keypoints have matches
  - Unnecessary to update these keypoints



Progressive matching and pose estimation
More accurate matches and precise pose

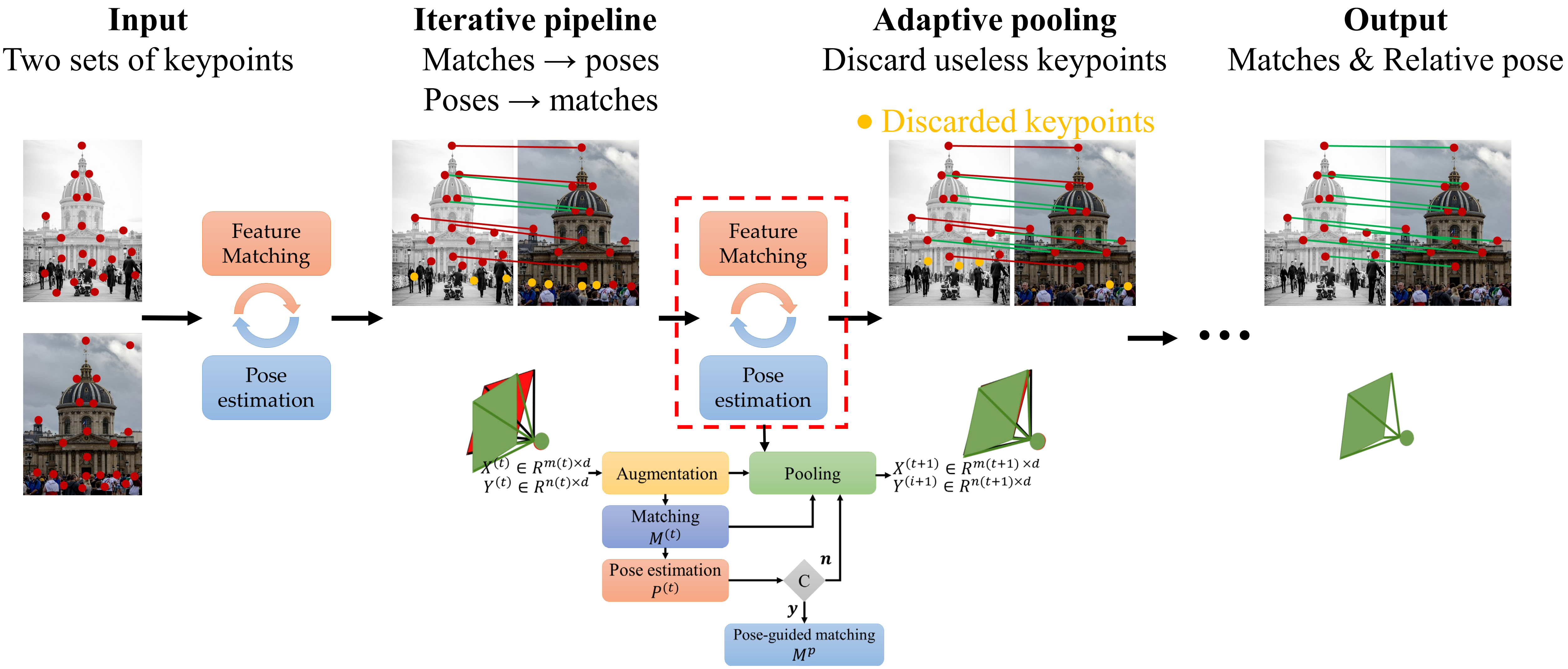


Detected keypoints

Keypoints with matches

- **Keypoints** 1024×1024
- **Matches** 285×285 – **27.8%**
- **Outliers** 739× 739 – **72.2%**

Iterative matching & pose estimation



Transformer-based recurrent module

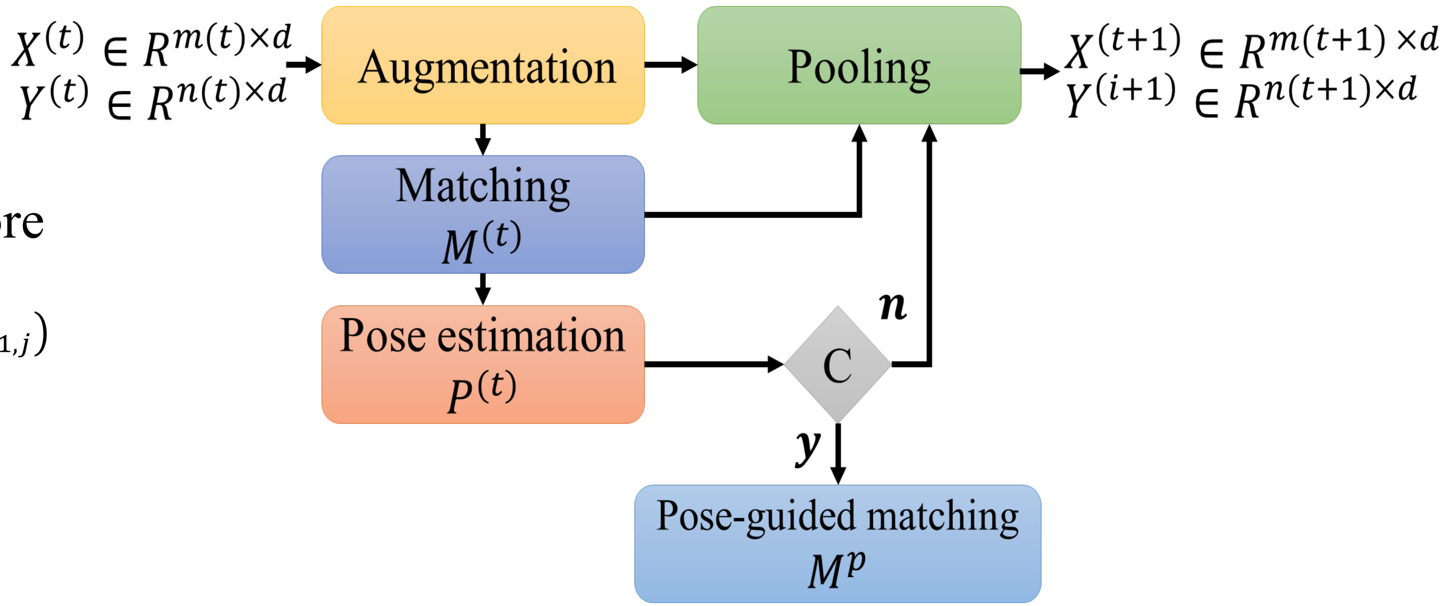
[1] Vaswani et al., Attention is all you need, NeurIPS 2017

Transformer-based recurrent module

1. **Transformer-based augmentation**
   - Descriptors augmented by spatial information
   - Quadratic complexity

Self attention **Cross attention**

$$X^{(t)} = X^{(t)} + f_A\left(X^{(i)}, X^{(i)}\right) + f_A\left(X^{(t)}, Y^{(t)}\right)$$
$$Y^{(t)} = Y^{(t)} + f_A\left(Y^{(t)}, X^{(t)}\right) + f_A\left(Y^{(t)}, Y^{(t)}\right)$$



2. **Cross entropy loss for matching**
   - Discriminative features have high score

$$L_M = -\sum_{(i,j)\in M} \log\left(\widehat{M}_{ij}\right) - \sum_i \log\left(\widehat{M}_{i,n+1}\right) - \sum_j \log\left(\widehat{M}_{m+1,j}\right)$$

3. **Pose-aware loss**
   - *Good* matches have *higher* score

$P = f_{w8}(x_j, y_j, M_{x_j y_j})$ **weighted 8pt pose estimation**

$$L_{pose} = l_2(P, P^{gt})$$

$$L_{geo} = \frac{1}{n} \frac{(y_i^T F x_i)^2}{||Fx_i||_{[1]}^2 + ||Fx_i||_{[2]}^2 + ||F^T y_i||_{[1]}^2 ||F^T y_i||_{[2]}^2}$$

Final loss

$$L_{final} = \alpha_M L_M + \alpha_{pose} L_{pose} + \alpha_{geo} L_{geo}$$

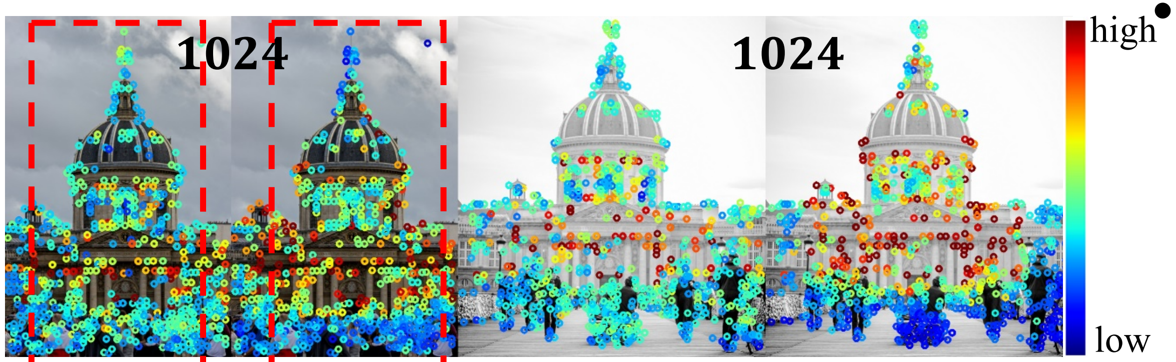
Correct matches

Pose-aware matches

[1] Sarlin et al., Superglue: Learning feature matching with graph neural networks, CVPR 2020
[2] Hartley and Zisserman, Multiple view geometry in computer vision, Cambridge university press 2003

Adaptive pooling

- **Attention score tells which are inliers**
  - keypoints with high scores ≈ inliers



Self and cross attention scores

Keypoints with potential correspondences

- **Our intention**
  - Keep as many inliers as possible
  - Remove as many low-contribution samples as possible

- **How to decide which one to discard**

Adaptive pooling

- **Using matching matrix as guidance**

Step 1: samples with high matching score as seeds (inliers)

$$\boldsymbol{X_M^{(t)}, Y_M^{(t)}, M_{X,Y} \geq \theta}$$

Samples (seeds) with potential matches

Step 2: retain samples with high attention scores with guidance (keypoints with high contribution)

Attention scores   Median value

$$X_A^{(t+1)} = X_{Self}^{(t)} \cup X_{Cross}^{(t)}, S(X_{Self/Corss}) \geq md(S(X_M^{(t)}))$$

$$Y_A^{(t+1)} = Y_{Self}^{(t)} \cup Y_{Cross}^{(t)}, S(Y_{Self/Corss}) \geq md(S(Y_M^{(t)}))$$

Finally kept keypoints

Step 3: merge samples with potential matches and high attention scores

$$X^{(t+1)} = X_M^{(t)} \cup X_A^{(t+1)}, Y^{(t+1)} = Y_M^{(t)} \cup Y_A^{(t+1)}$$

Number of keypoints: **1024 -> 496/385**

Adaptive pooling

- **Uncertainty-aware pooling**
  - Matches could be wrong due to large viewpoint changes
  - Poses reveal the quality of matches

Step 2: retain samples with high attention scores with guidance

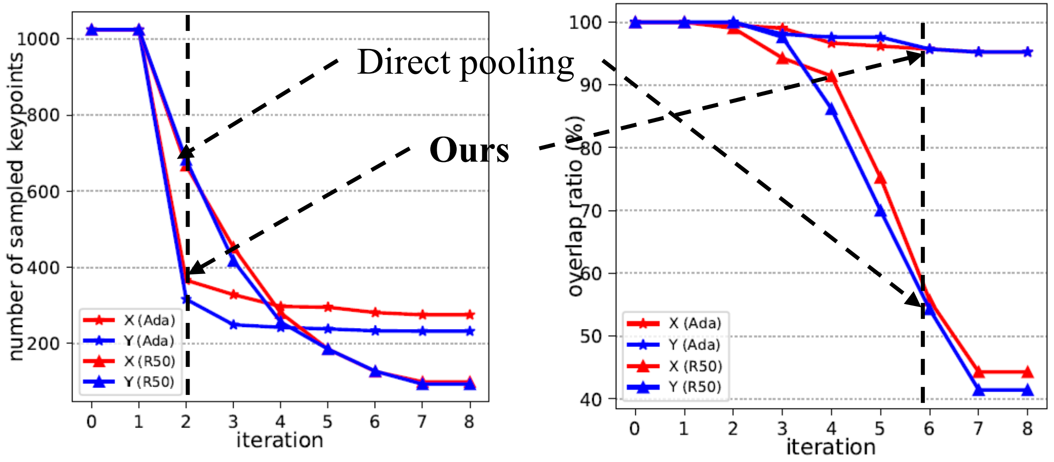
Attention scores Median value

$$X_A^{(t)} = X_{Self}^{(t)} \cup X_{Cross}^{(t)}, S(X_{Self/Corss}) \geq md(S(X_M^{(t)})) * \tau$$

$$Y_A^{(t)} = Y_{Self}^{(t)} \cup Y_{Cross}^{(t)}, S(Y_{Self/Corss}) \geq md(S(Y_M^{(t)})) * \tau$$

$$\tau = \frac{|(x_i, y_i), s.t., f_{epipolar}(x_i, y_i, P^t) \leq \theta_{epipolar}|}{|(x_i, y_i) \in M^{(t)}|}$$

Effective outlier removing Effective inlier preserving



Preserved keypoints and ratio of inliers

Pose not accurate → matches not good → keep more samples
Pose accurate → matches good → keep fewer samples

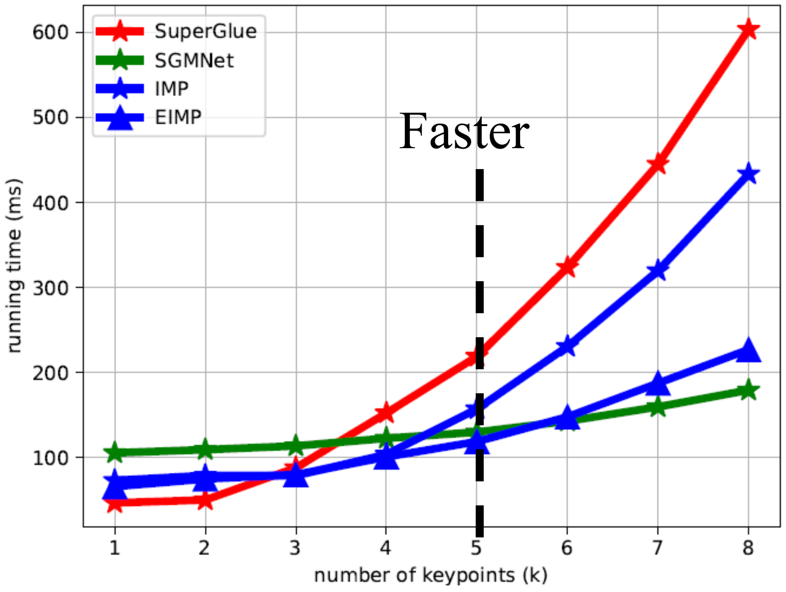
Quantitative results

- **Training**
  - Megadepth dataset from scratch without any pretraining
- **Better pose accuracy**
  - Outdoor YFCC and Indoor Scannet datasets

Group	Method	@5	@10	@20	@5	@10	@20
	NN-mutual	6.5	15.4	28.5	9.4	21.6	36.4
Filtering	AdaLAM	20.8	36.5	51.9	6.7	15.8	27.4
	OANet	19.2	34.5	50.3	10.0	25.1	38.0
	CLNet	27.8	46.4	63.8	4.1	11.0	21.6
Graph-matcher	SuperGlue	37.1	57.2	73.6	16.2	32.6	49.3
	SGMNet	35.3	56.1	73.6	16.4	32.1	48.7
	IMP	**39.4**	**59.4**	**75.2**	**16.6**	**33.1**	**49.4**
	EIMP	37.9	57.9	74.0	15.9	32.4	48.9

Relative pose accuracy on YFCC and Scannet datasets
The **best** and second-best are highlighted.

- **Higher speed**
  - IMP is faster than SuperGlue
  - EIMP is close to SGMNet

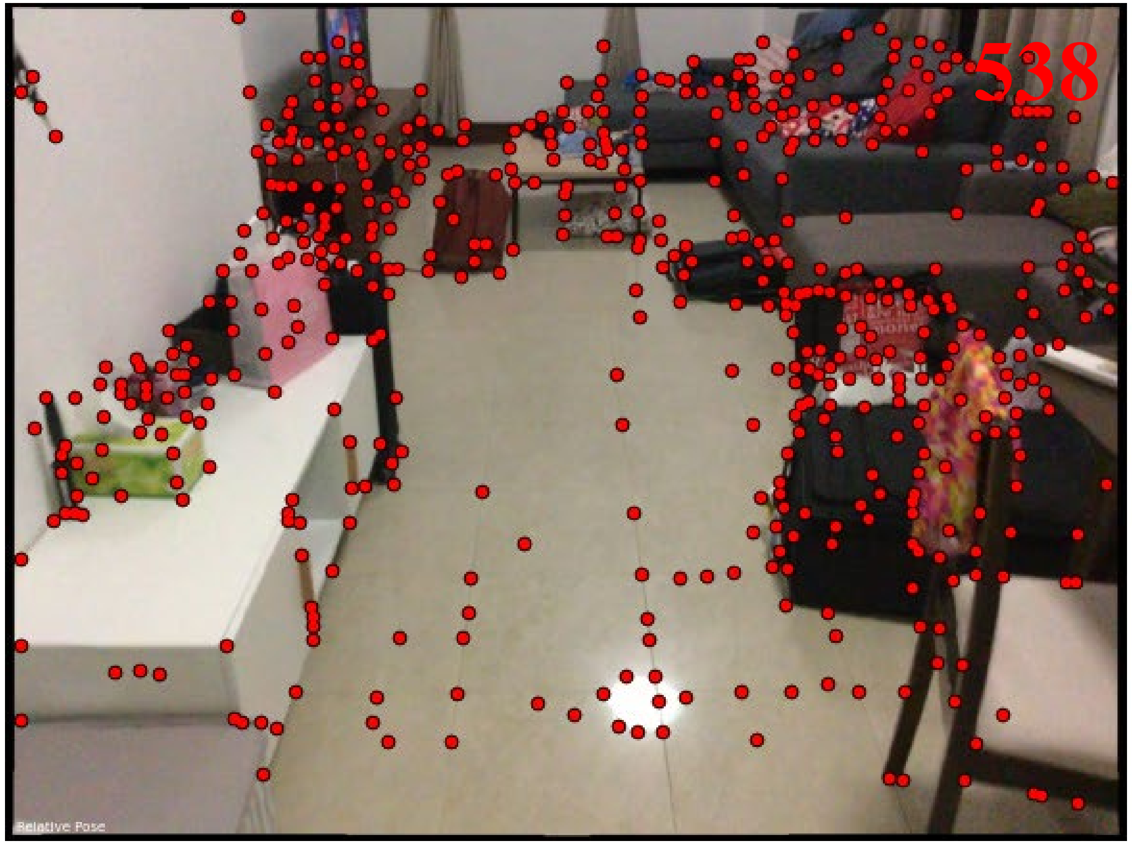


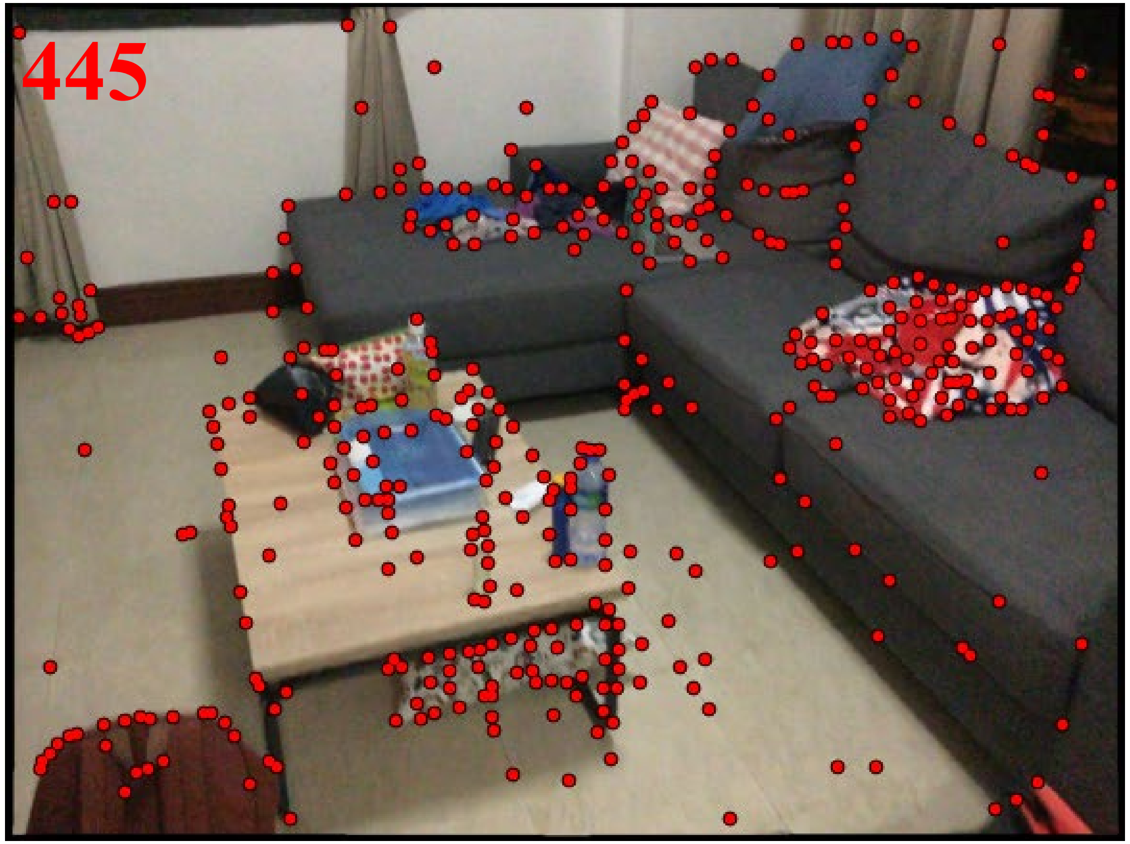
Running time of different #keypoints

[1] Zhang et al., Learning two-view correspondences and geometry using order-aware network, ICCV 2019
[2] Sarlin et al., Superglue: Learning feature matching with graph neural networks, CVPR 2020
[3] Li and Snavely, Megadepth: Learning singleview depth prediction from internet photos. CVPR 2018
[4] Thomee et al., YFCC100M: The new data in multimedia research, Communications of the ACM 2016
[5] Dai et al., Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration, ACM ToG 2017
[6] Zhao et al., Progressive correspondence pruning by consensus learning, ICCV 2021
[7] Chen et al., Learning to match features with seeded graph matching network, CVPR 2021

Results on Scannet dataset - case 1

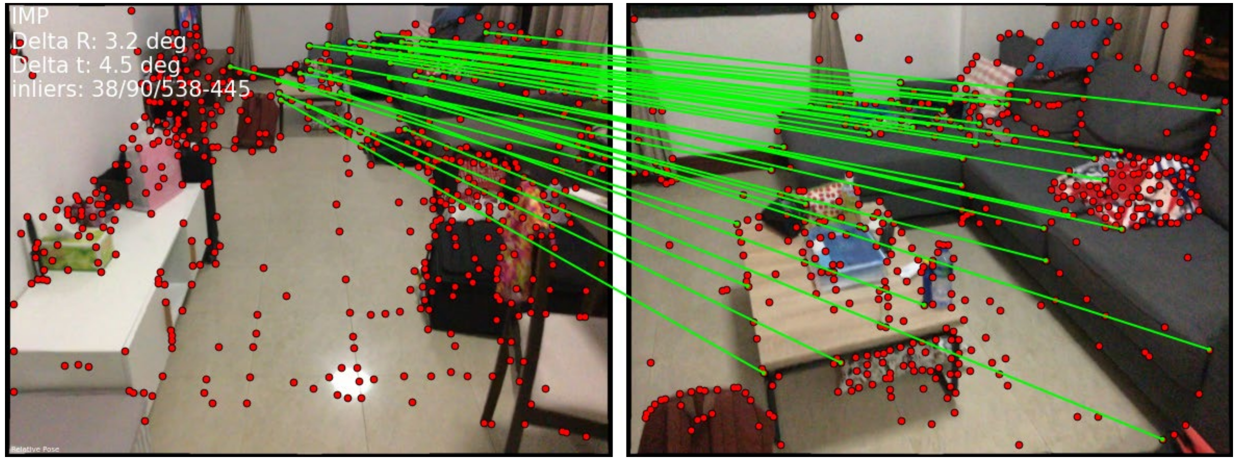
Extracted keypoints





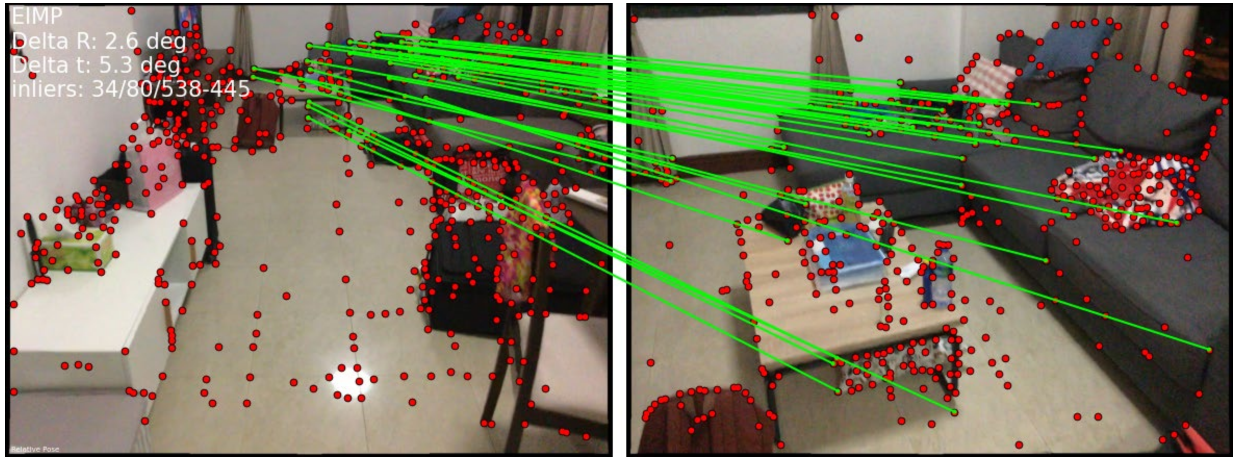
Results on Scannet dataset - case 1

Inliers/matches: 38/96, R/t error: 3.2/4.5deg
Keypoints left/right: 538/445



IMP (iteration 1)

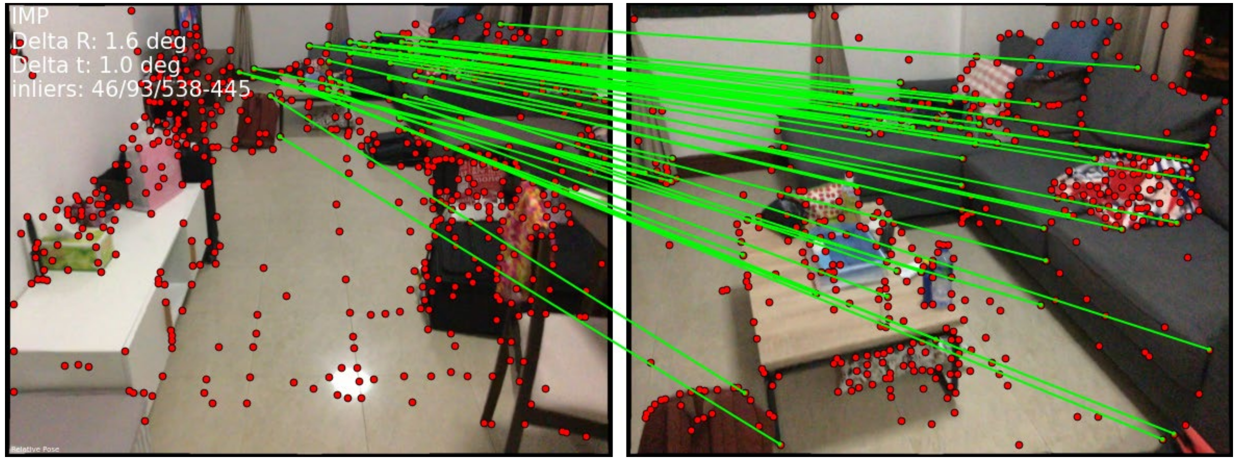
Inliers/matches: 34/80, R/t error: 2.6/5.3deg
Keypoints left/right: 538/445



EIMP (iteration 1)

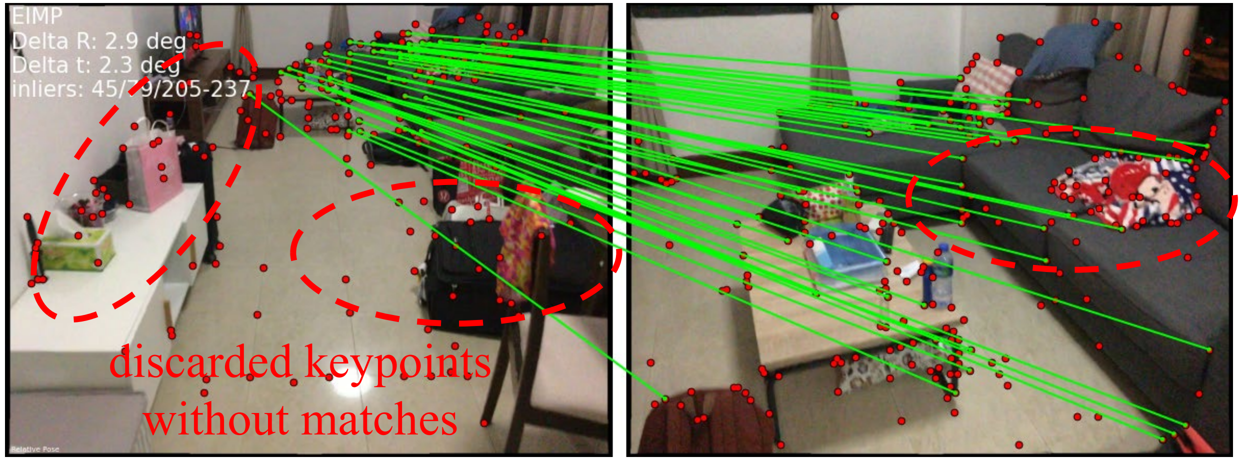
Results on Scannet dataset - case 1

Inliers/matches: 46/93, R/t error: 1.6/1.0deg
Keypoints left/right: 538/445



IMP (iteration 2)

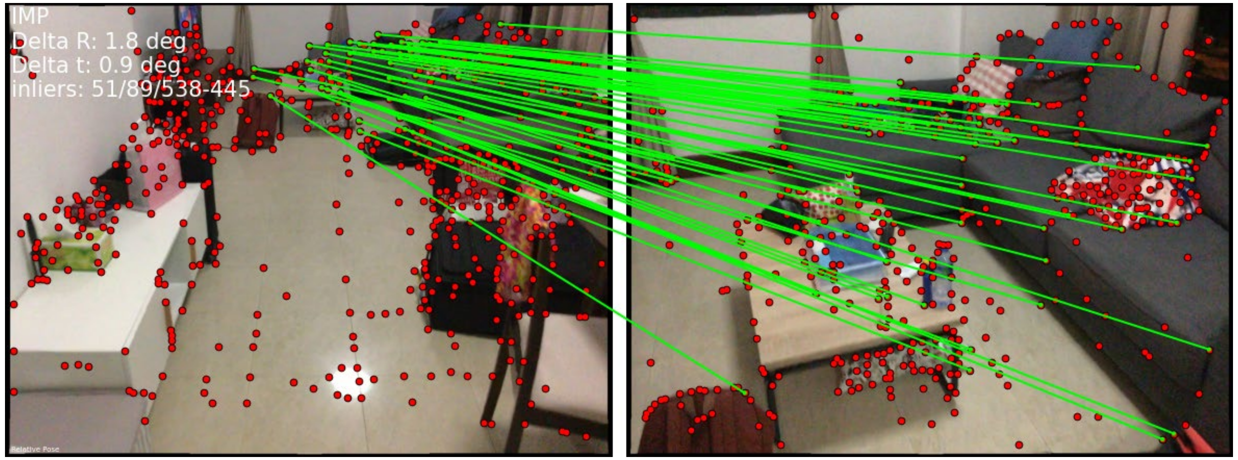
Inliers/matches: 45/79, R/t error: 2.9/2.3deg
Keypoints left/right: 205/237



EIMP (iteration 2)

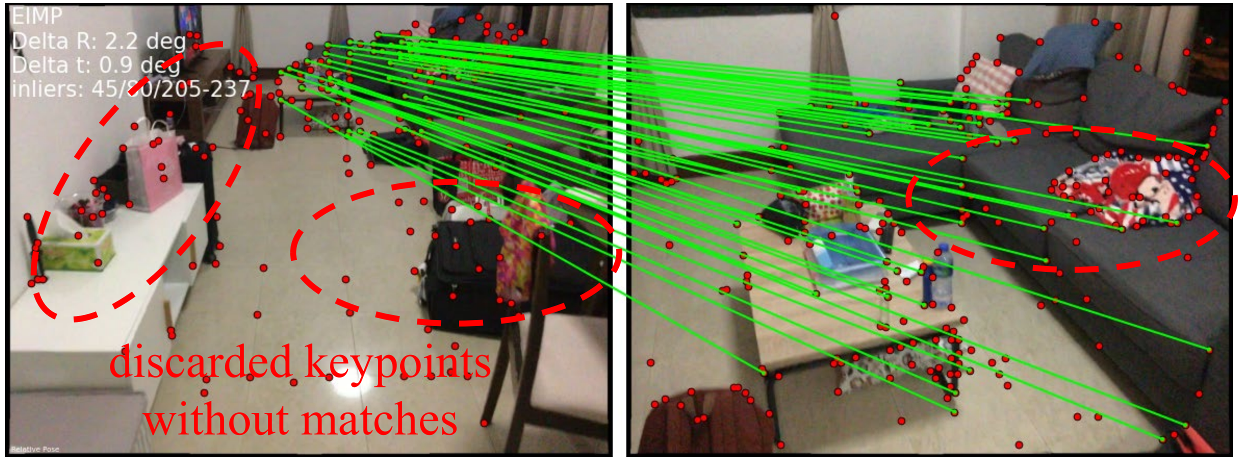
Results on Scannet dataset - case 1

Inliers/matches: 51/89, R/t error: 1.8/0.9deg
Keypoints left/right: 538/445



IMP (iteration 3)

Inliers/matches: 45/80, R/t error: 2.2/0.9deg
Keypoints left/right: 205/237



EIMP (iteration 3)

Results on Scannet dataset - case 1

IMP

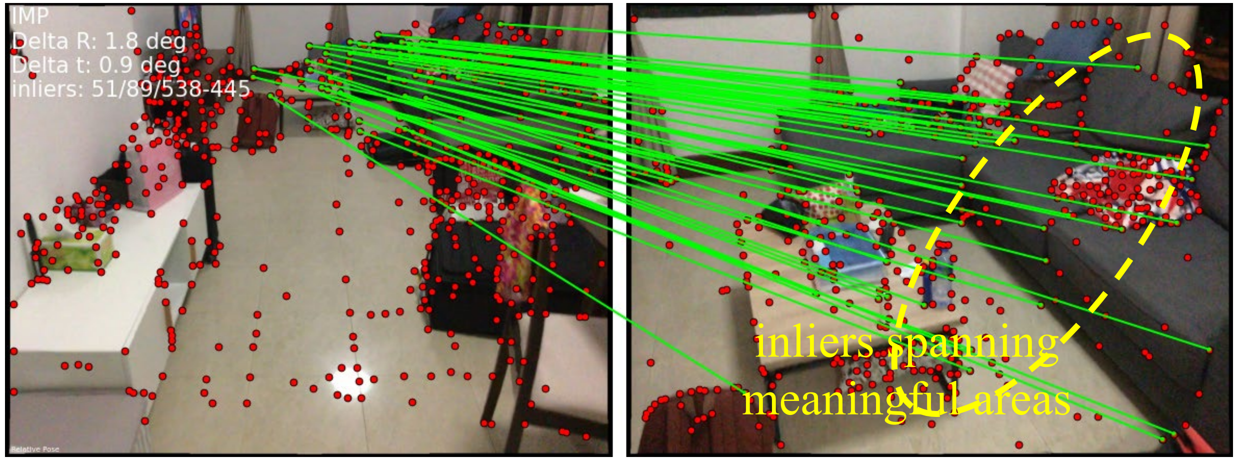
Inliers/matches: 46/93, R/t error: 1.6/1.0deg

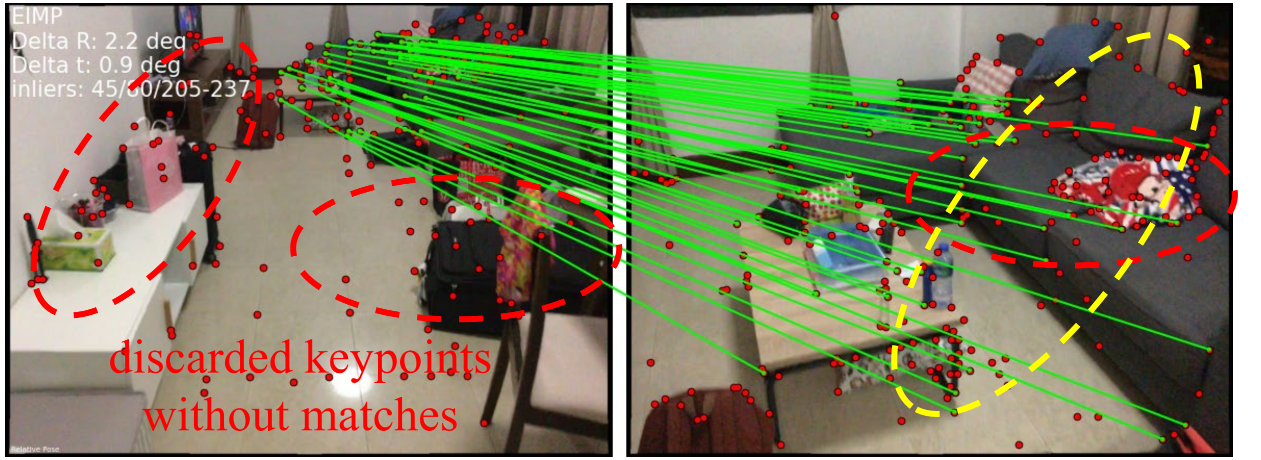
Keypoints left/right: 538/445

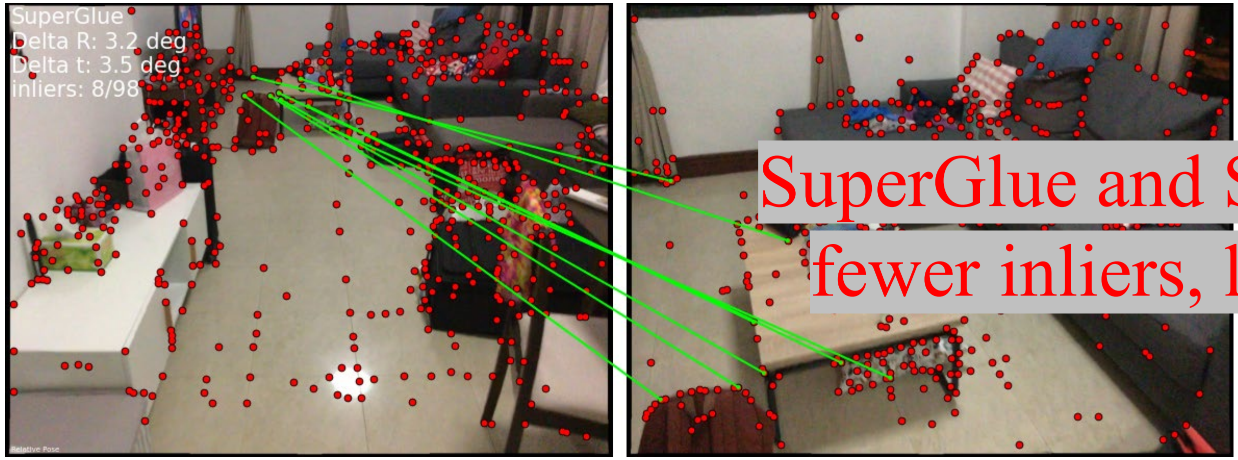
EIMP

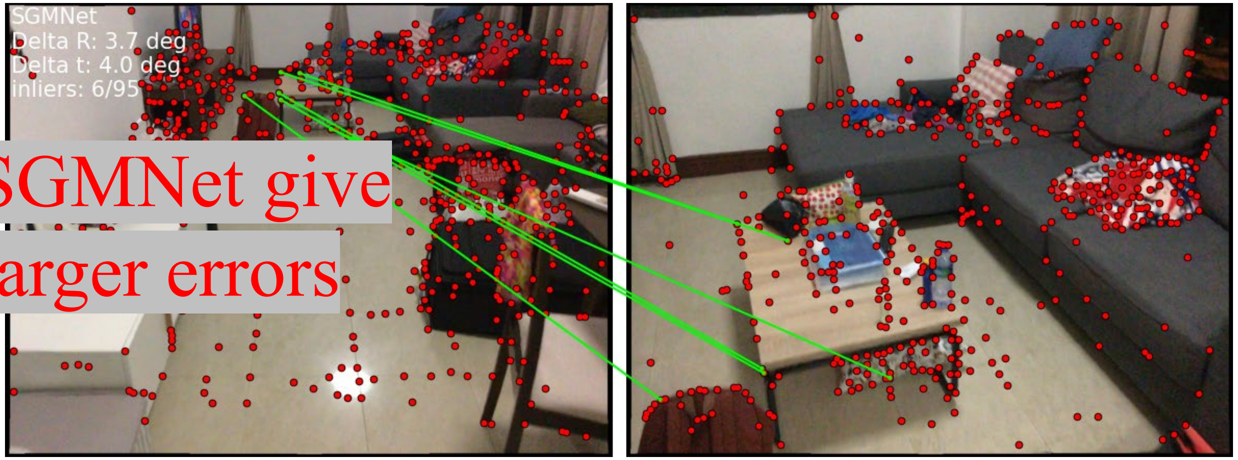
Inliers/matches: 45/79, R/t error: 2.9/2.3deg

Keypoints left/right: 205/237









SuperGlue and SGMNet give fewer inliers, larger errors

Inliers/matches: 8/98, R/t error: 3.2/3.5deg

Keypoints left/right: 538/445

SuperGlue

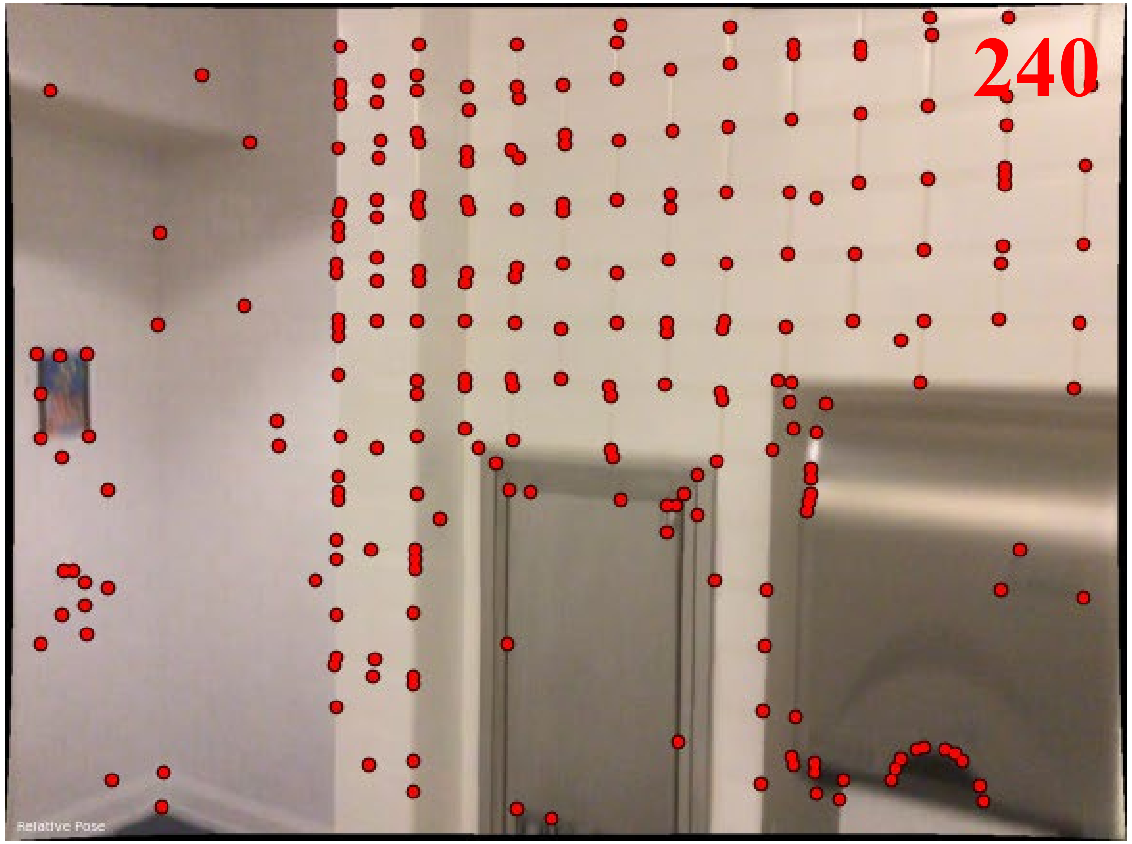
Inliers/matches: 6/95, R/t error: 3.7/4.0deg

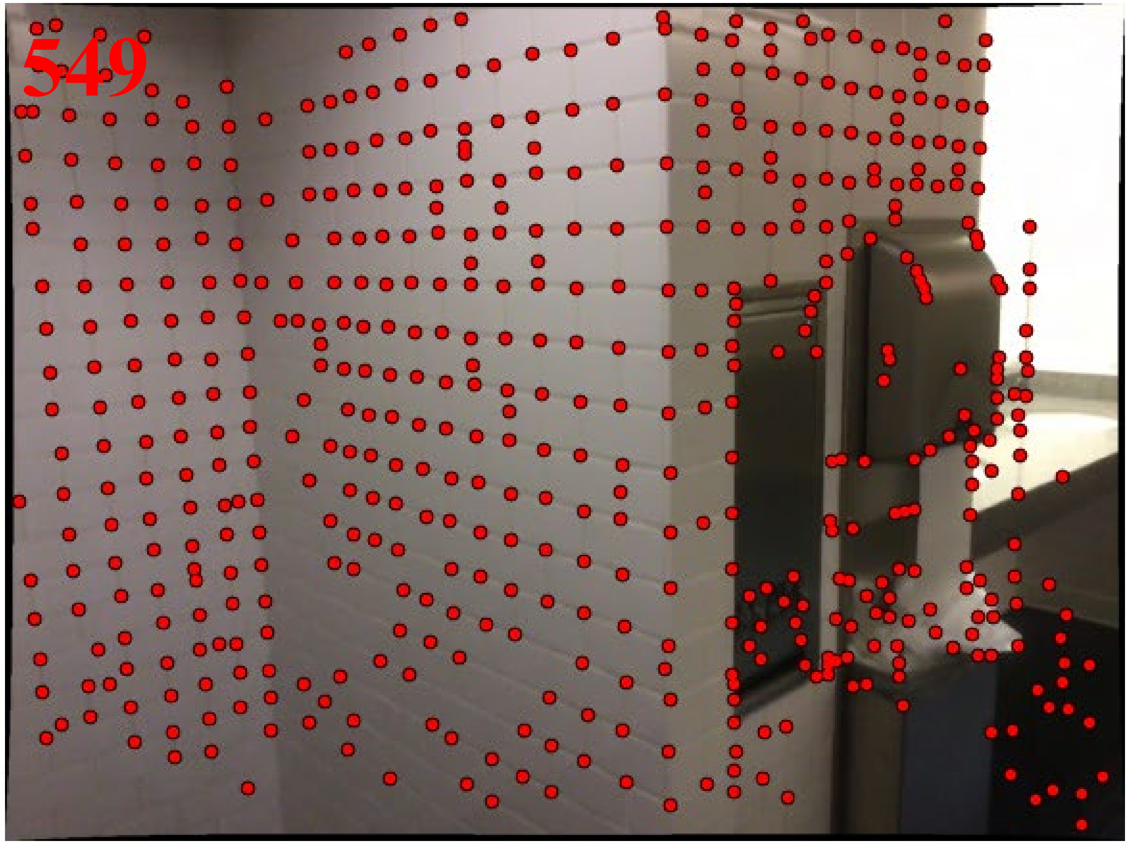
Keypoints left/right: 538/445

SGMNet

Results on Scannet dataset - case 2

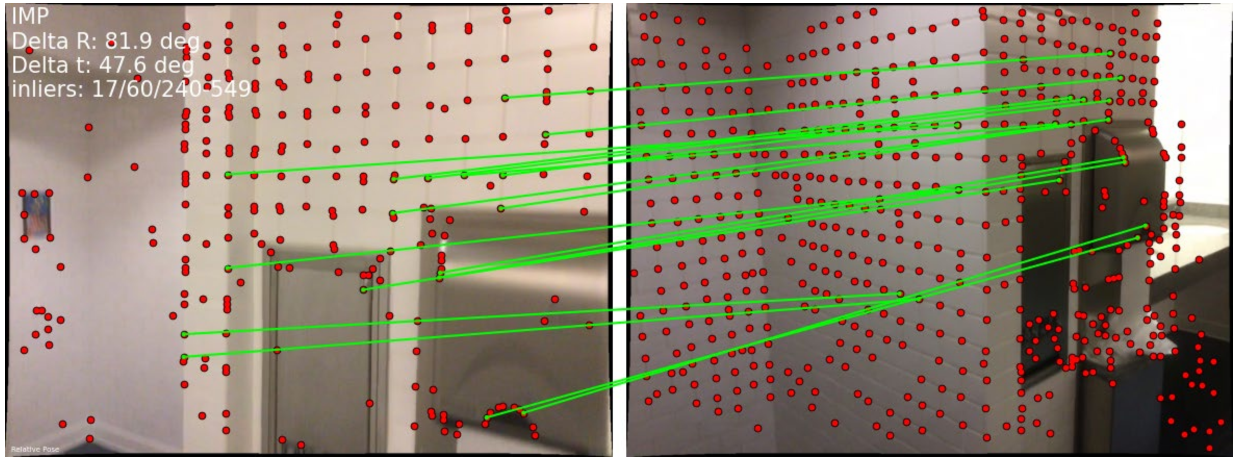
Extracted keypoints





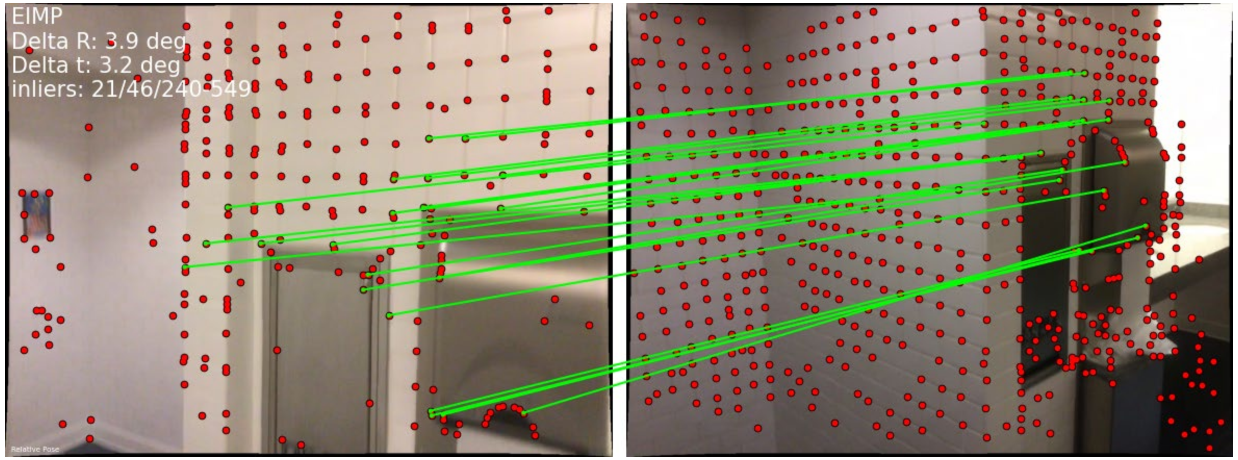
Results on Scannet dataset - case 2

Inliers/matches: 17/60, R/t error: 81.9/47.6deg
Keypoints left/right: 240/549



IMP (iteration 1)

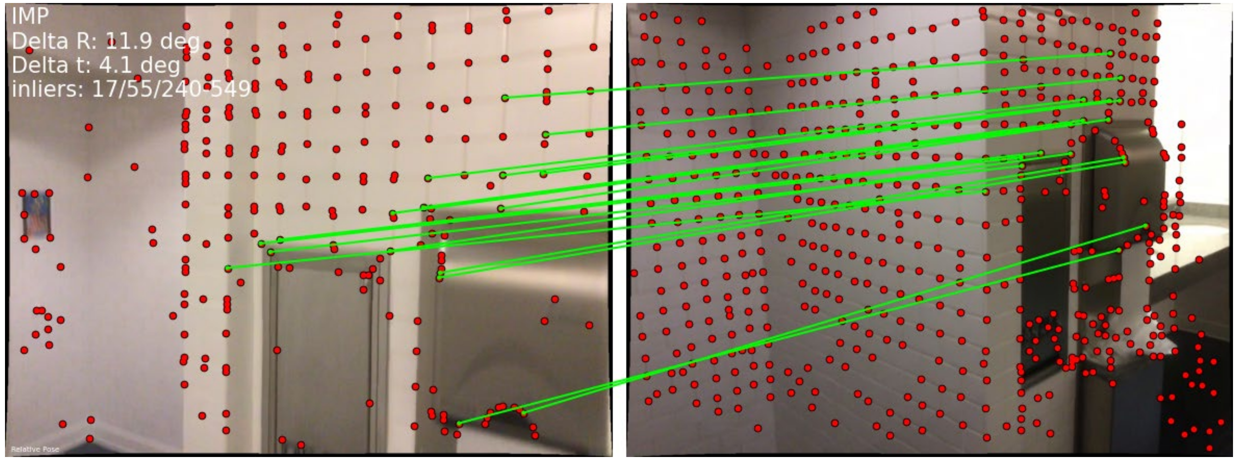
Inliers/matches: 21/46, R/t error: 3.9/3.2deg
Keypoints left/right: 240/549



EIMP (iteration 1)

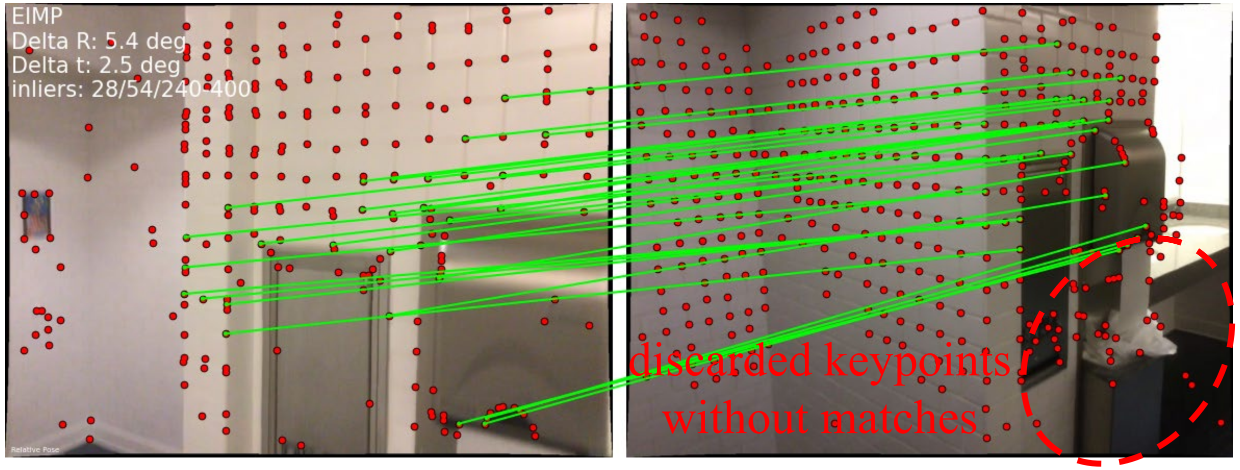
Results on Scannet dataset - case 2

Inliers/matches: 17/55, R/t error: 11.9/4.1deg
Keypoints left/right: 240/549



IMP (iteration 2)

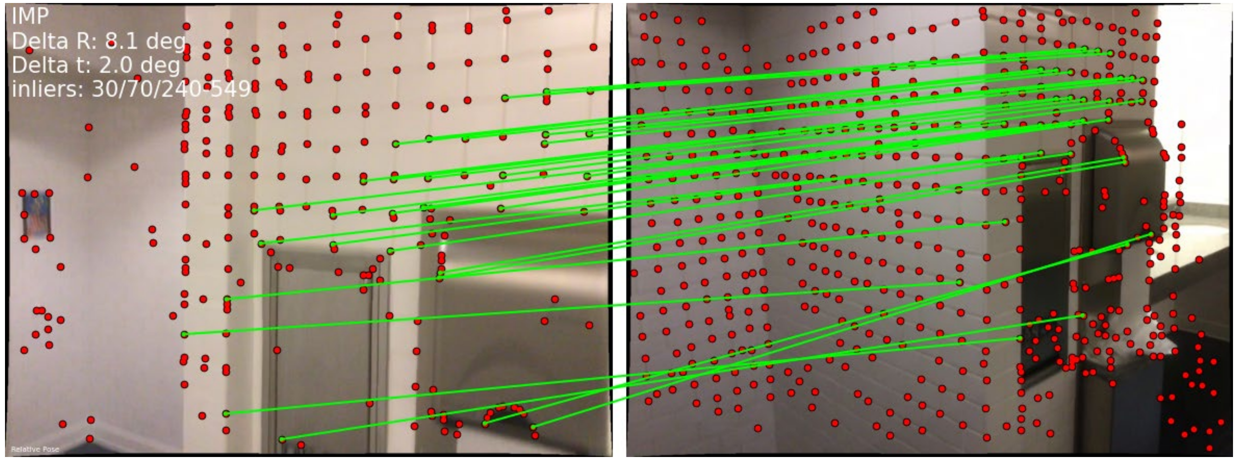
Inliers/matches: 28/54, R/t error: 5.4/2.5deg
Keypoints left/right: 240/400



EIMP (iteration 2)

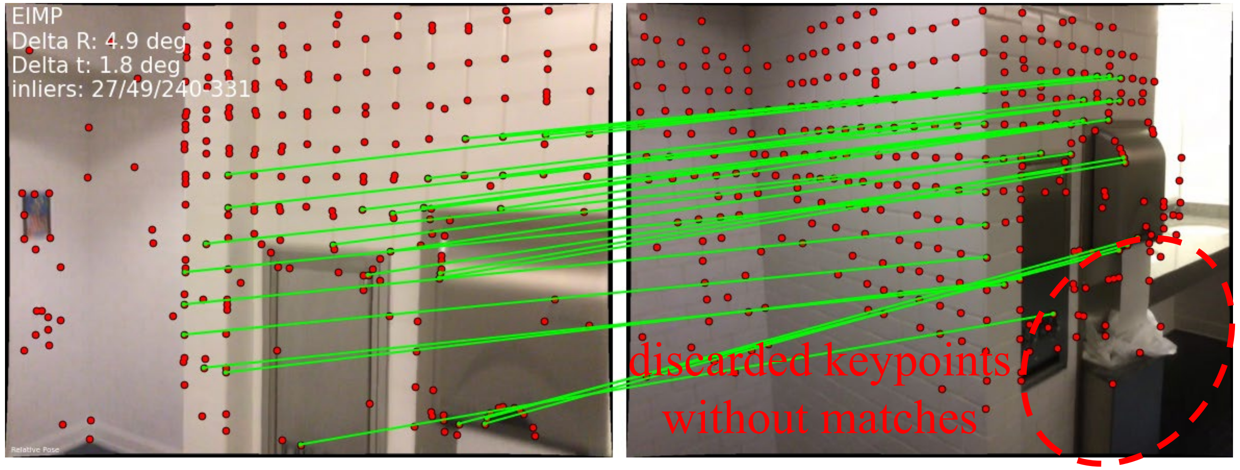
Results on Scannet dataset - case 2

Inliers/matches: 30/70, R/t error: 8.1/2.0deg
Keypoints left/right: 240/549



IMP (iteration 3)

Inliers/matches: 27/49, R/t error: 4.9/1.8deg
Keypoints left/right: 240/381



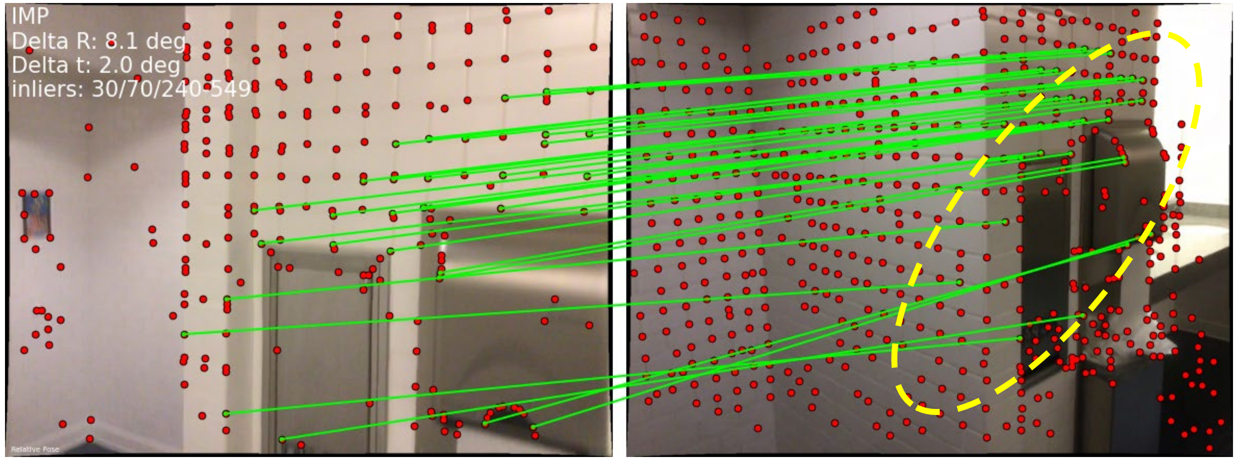
EIMP (iteration 3)

Results on Scannet dataset - case 2

IMP

Inliers/matches: 30/70, R/t error: 8.1/2.0deg

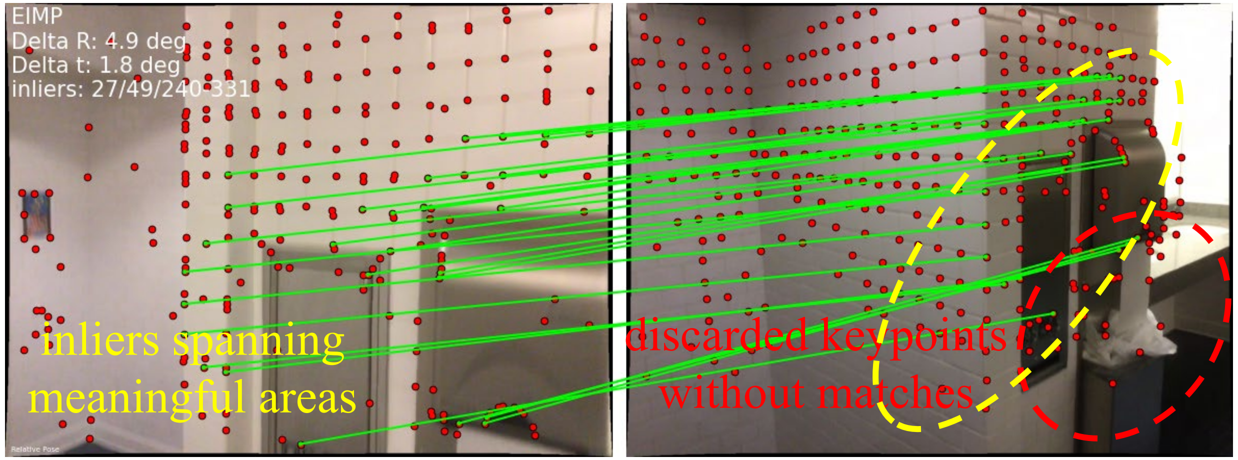
Keypoints left/right: 240/549



EIMP

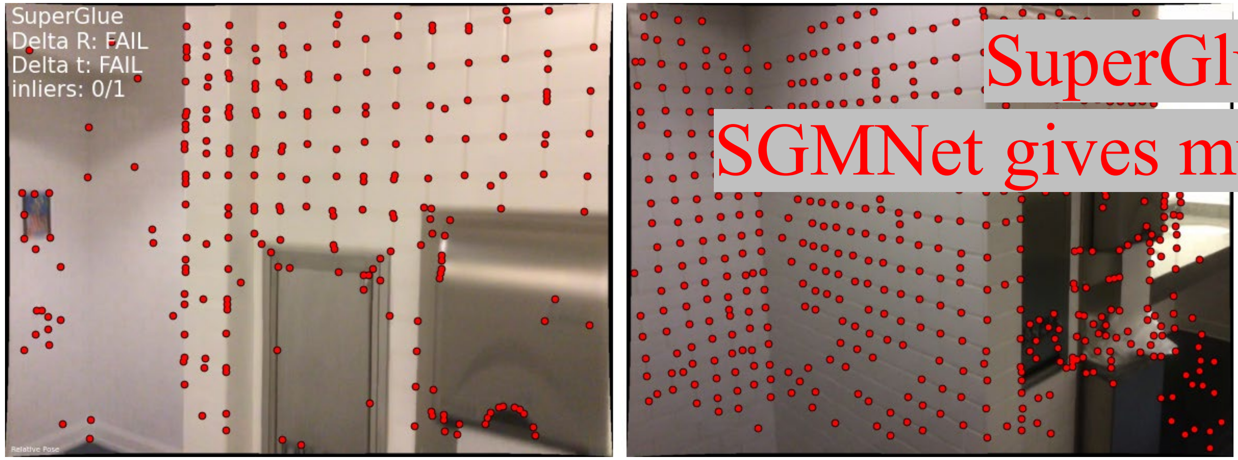
Inliers/matches: 27/49, R/t error: 4.9/1.8deg

Keypoints left/right: 240/381



SuperGlue fails,

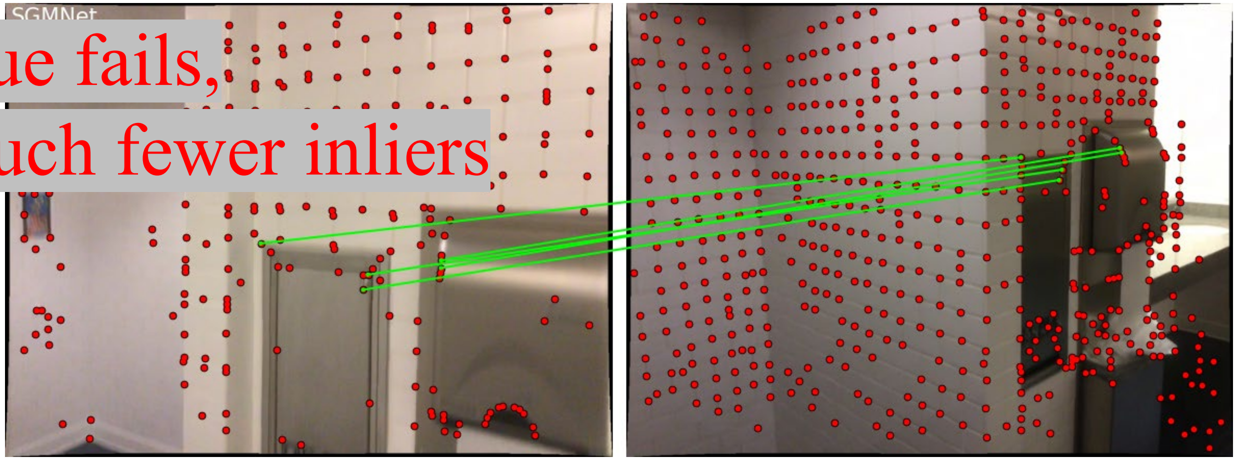
SGMNet gives much fewer inliers



Inliers/matches: 0/1, R/t error: FAIL

Keypoints left/right: 240/549

SuperGlue



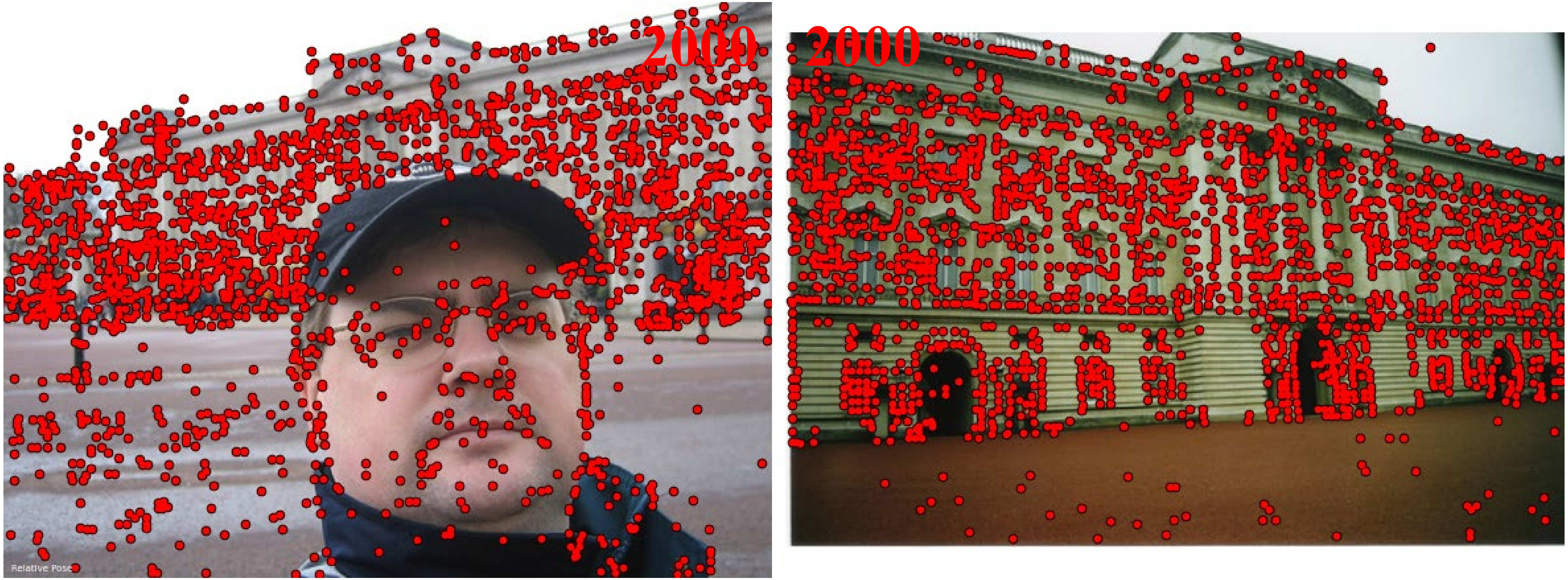
Inliers/matches: 5/41, R/t error: 16.1/8.1deg

Keypoints left/right: 240/549

SGMNet

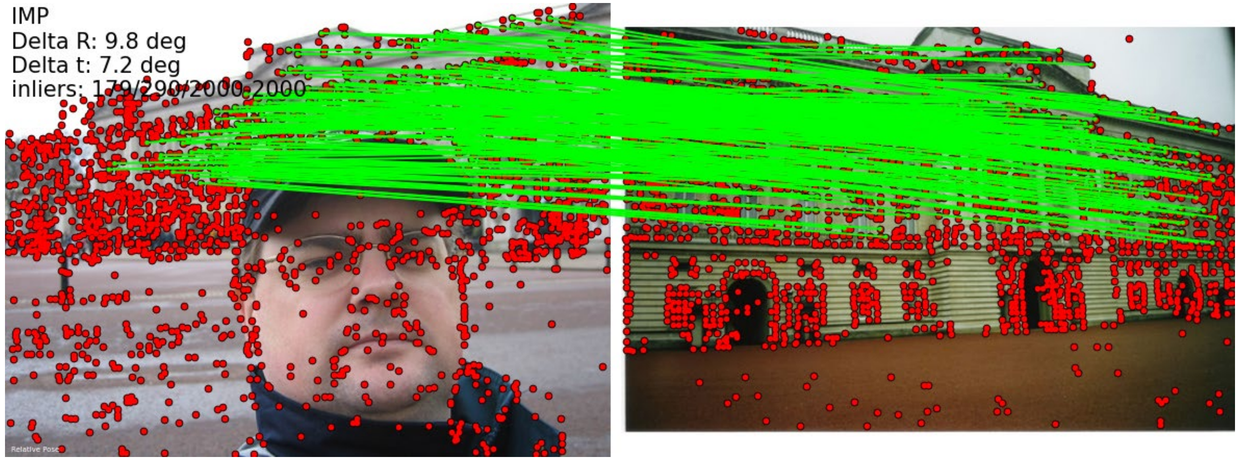
Results on YFCC100m dataset - case 1

Extracted keypoints



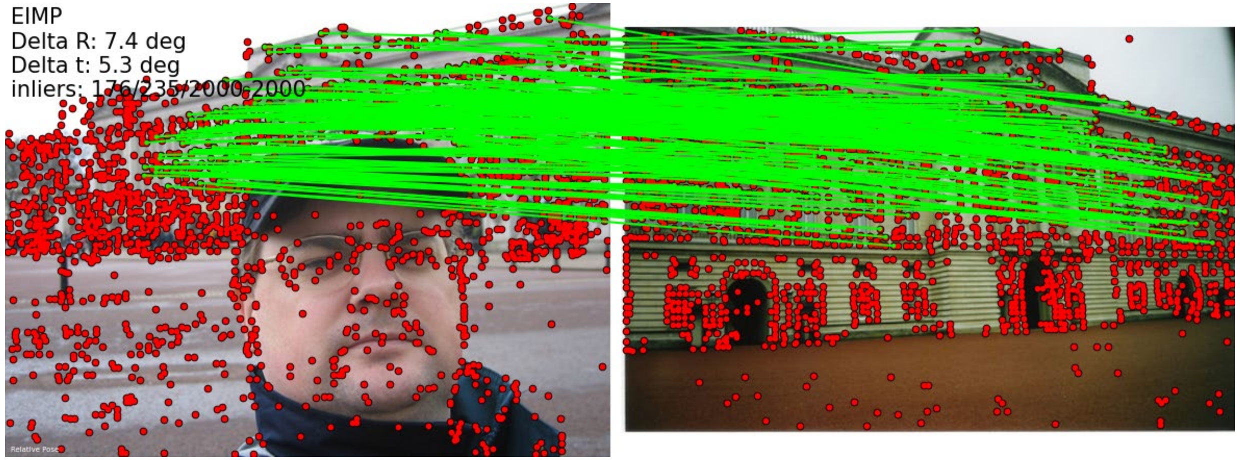
Results on YFCC100m dataset - case 1

Inliers/matches: 179/290, R/t error: 9.8/7.2deg
Keypoints left/right: 2000/2000



IMP (iteration 1)

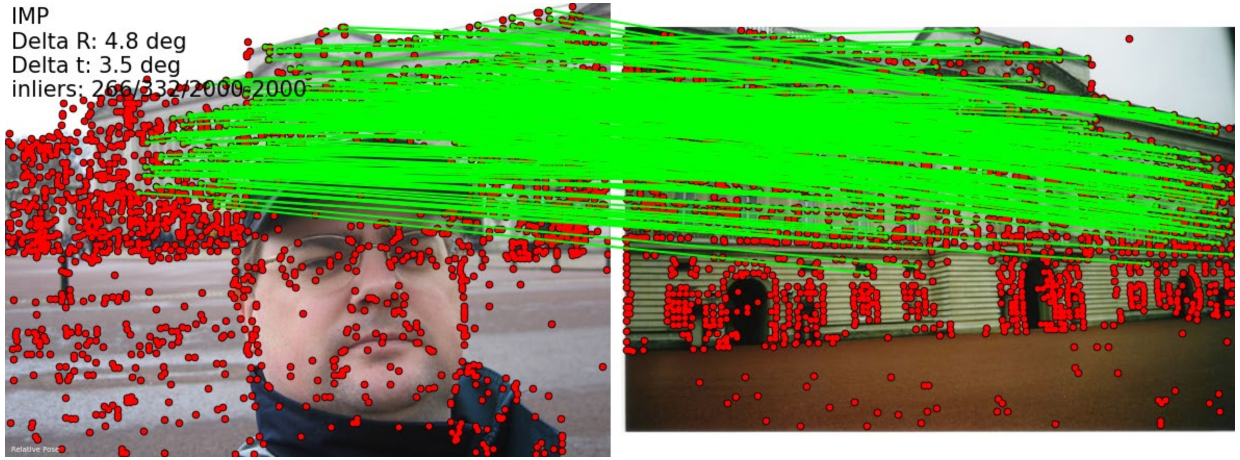
Inliers/matches: 126/235, R/t error: 7.4/5.3deg
Keypoints left/right: 2000/2000



EIMP (iteration 1)

Results on YFCC100m dataset - case 1

Inliers/matches: 266/332, R/t error: 4.8/3.5deg
Keypoints left/right: 2000/2000



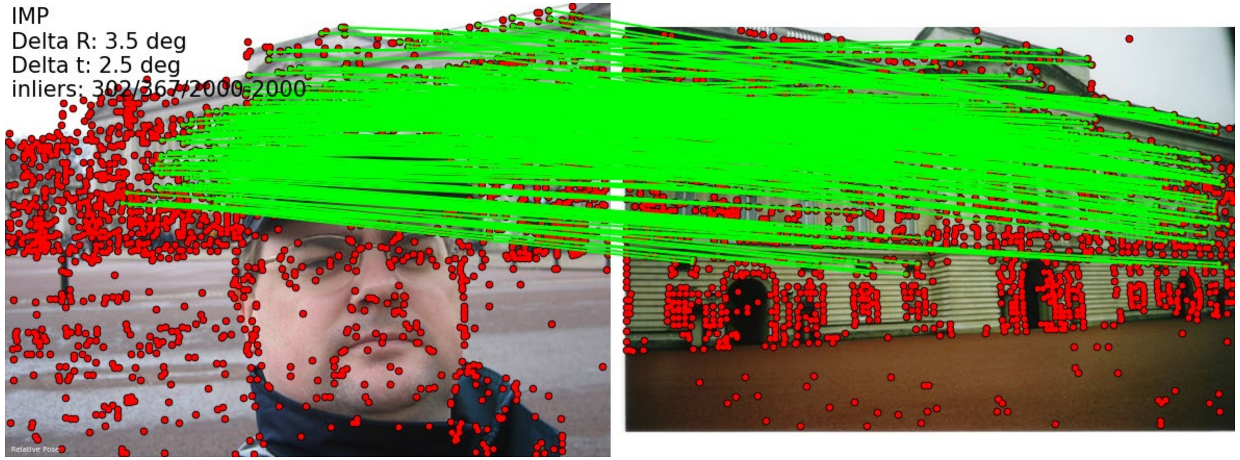
IMP (iteration 2)

Inliers/matches: 262/357, R/t error: 5.8/4.4deg
Keypoints left/right: 1167/1284

EIMP (iteration 2)

Results on YFCC100m dataset - case 1

Inliers/matches: 302/367, R/t error: 3.5/2.5deg
Keypoints left/right: 2000/2000



IMP (iteration 3)

Inliers/matches: 274/293, R/t error: 4.2/3.1deg
Keypoints left/right: 600/677

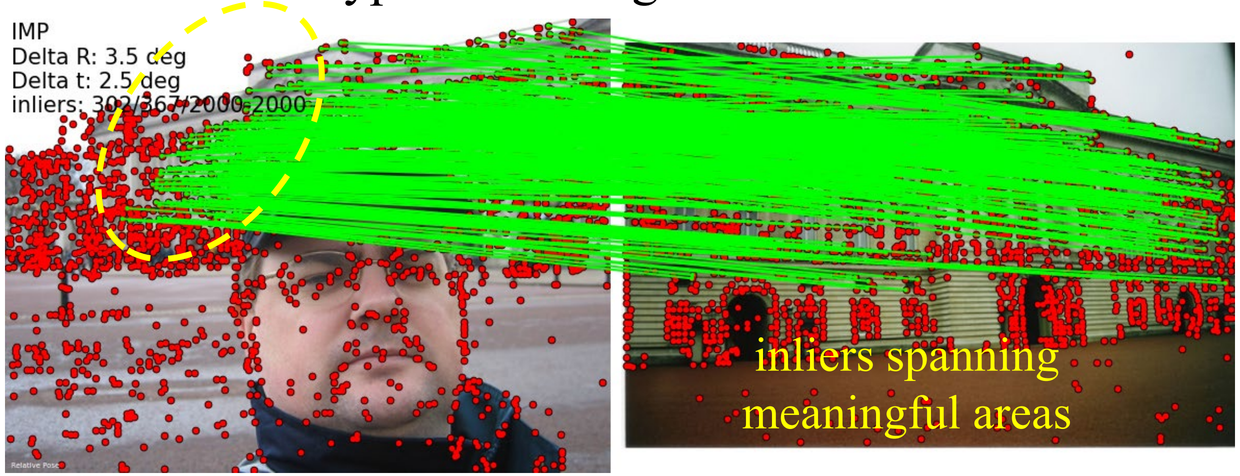
EIMP (iteration 3)

Results on YFCC100m dataset - case 1

IMP

Inliers/matches: 302/367, R/t error: 3.5/2.5deg

Keypoints left/right: 2000/2000



EIMP

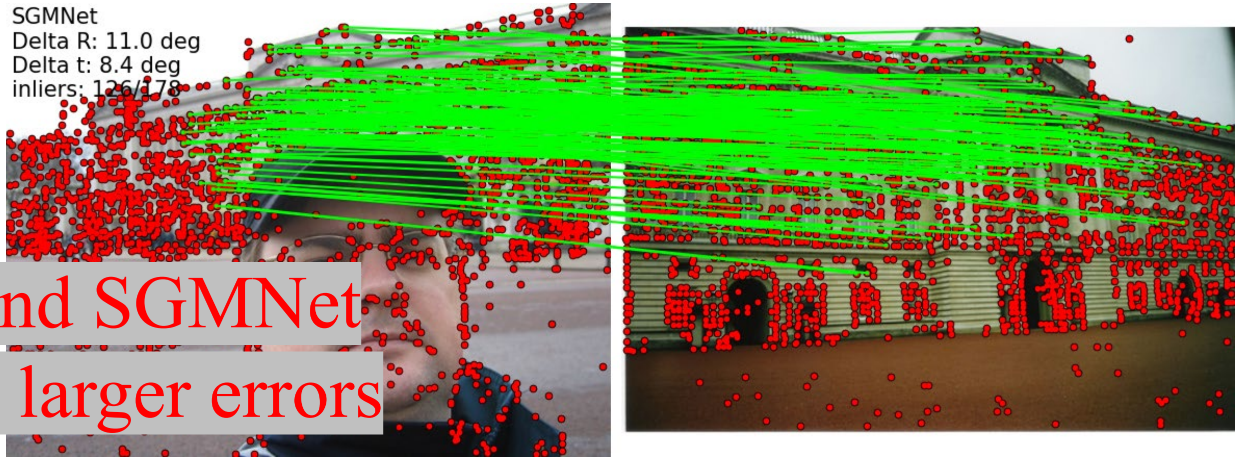
Inliers/matches: 274/293, R/t error: 4.2/3.1deg

Keypoints left/right: 600/677

Inliers/matches: 21/73, R/t error: 11.7/8.9deg

Keypoints left/right: 2000/2000

SuperGlue



Inliers/matches: 126/178, R/t error: 11.0/8.4deg

Keypoints left/right: 2000/2000

SGMNet

Conclusion and future work

- **Iterative matching and pose estimation**
  - Finding matches and estimating poses iteratively
  - Discarding useless keypoints dynamically

- **Future work**
  - Replacing traditional pose estimation with deep models