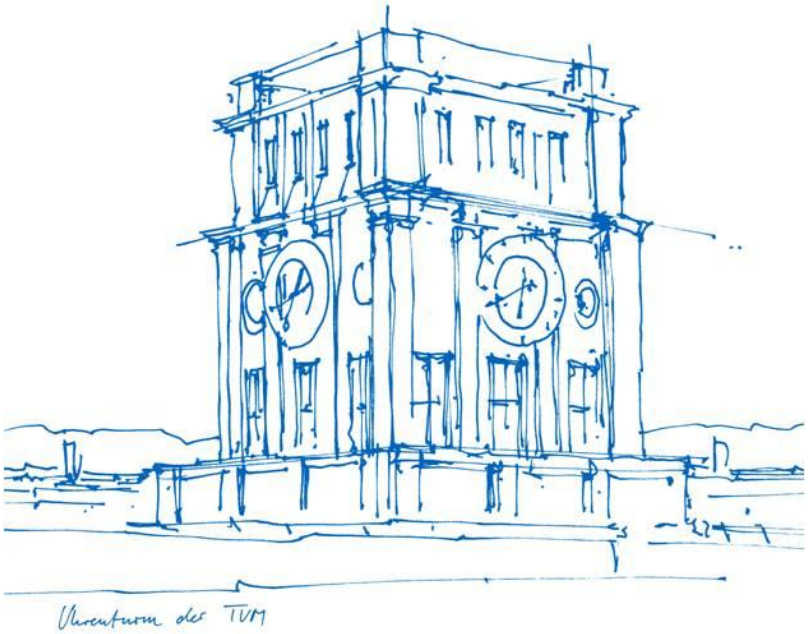
Learning Correspondences For Relative Pose Estimation

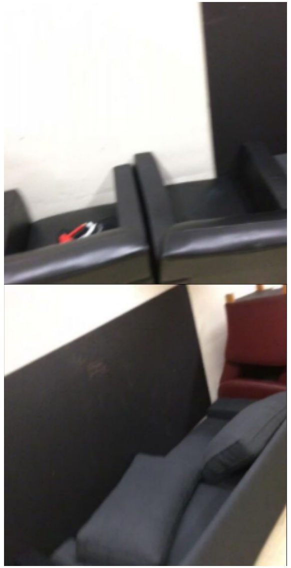
Guided Research

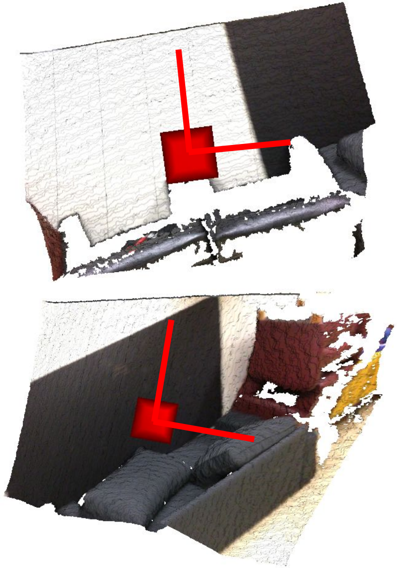


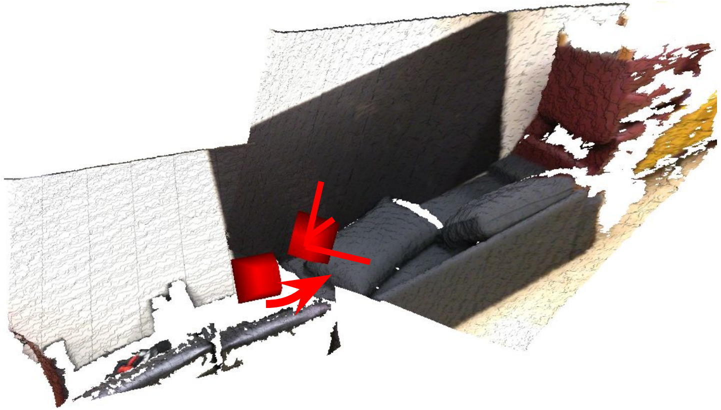
Marc Benedí San Millán
Supervisor: Prof. Dr. Matthias Nießner

Problem

- Input is a pair of RGB-D frames
- Estimate relative camera position

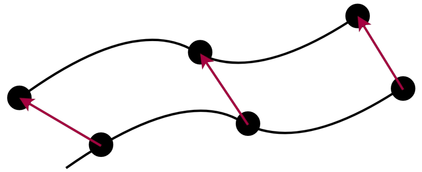


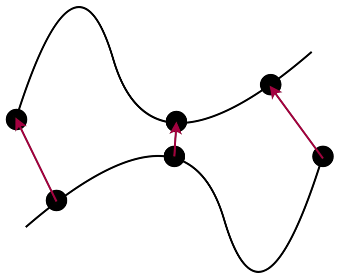




Motivation

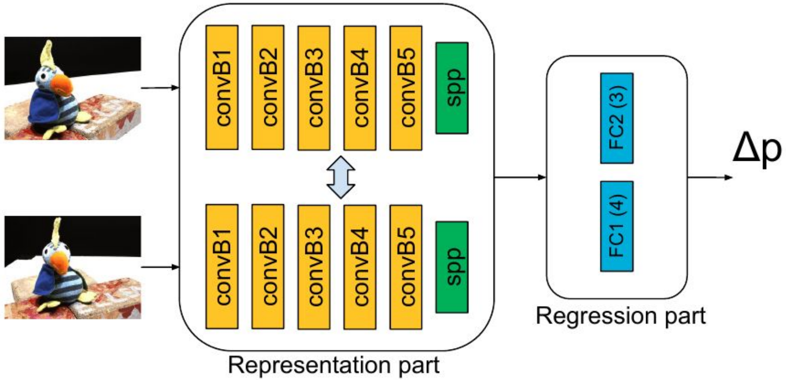
- Iterative Closest Point (ICP) [Besl & McKay, 92]
  - Point cloud alignment
  - Two steps
    - Data association
    - Transformation estimation
  - Converges to a good alignment if starting positions are ‘close enough’
    - Problem: it doesn’t converge otherwise
- Model that provides initial alignment





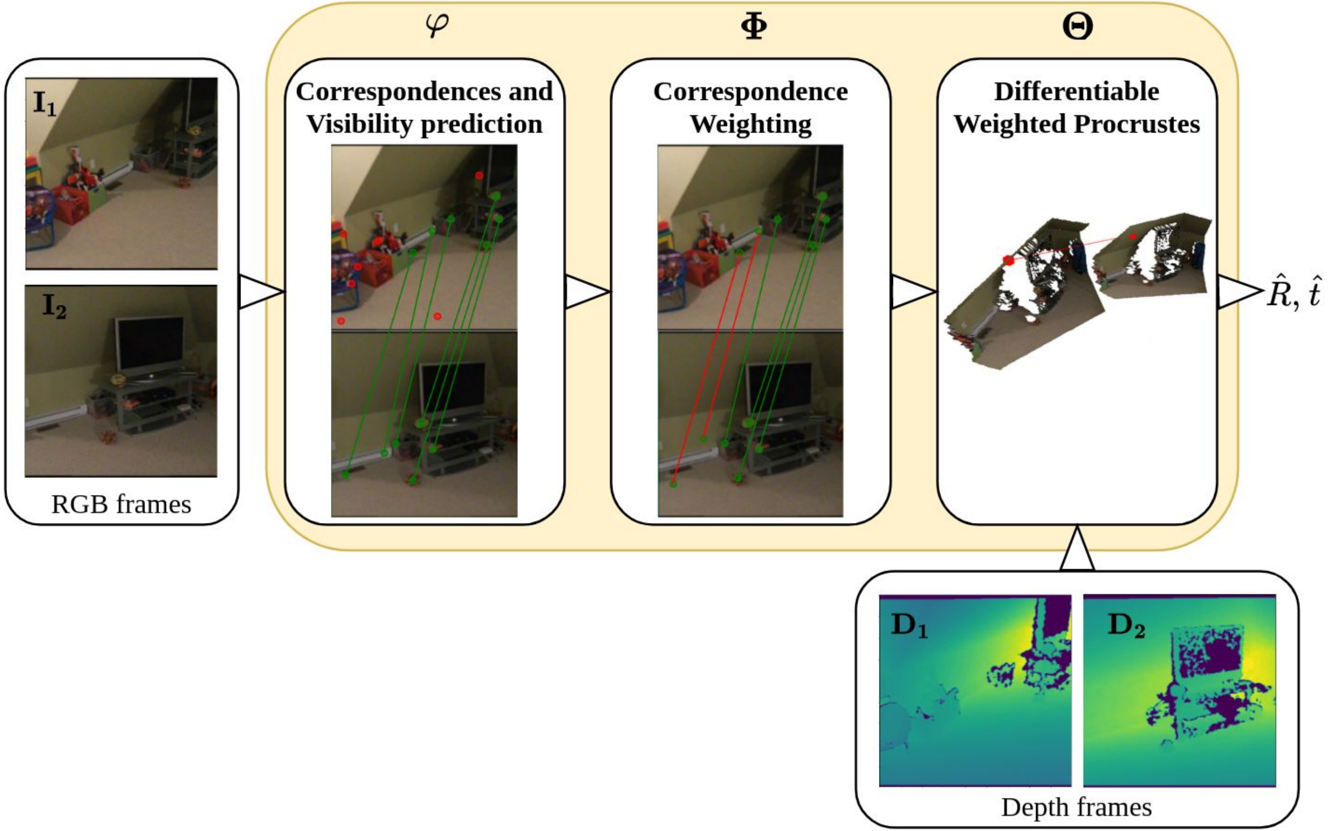
Direct Regression approaches

- PoseNet [Kendall et al, 15]
  - Use CNN encoder and FC regresor to estimate the absolute pose
  - Multiple approaches extended this idea for relative pose estimation

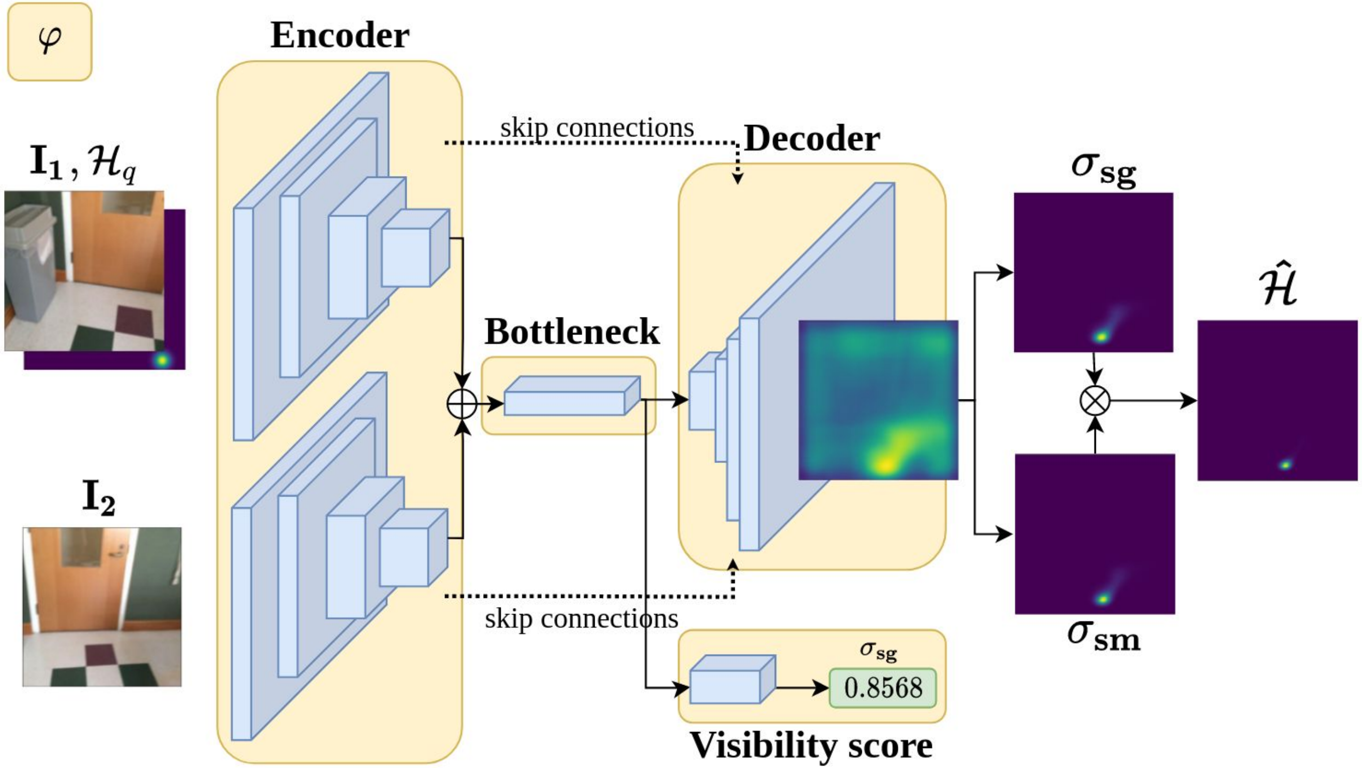


RelPoseNet[Melekhov et al, 17]

Method - Overview



Method - Correspondence and Visibility Predictor



DeepDeform [Božič et al, 19]

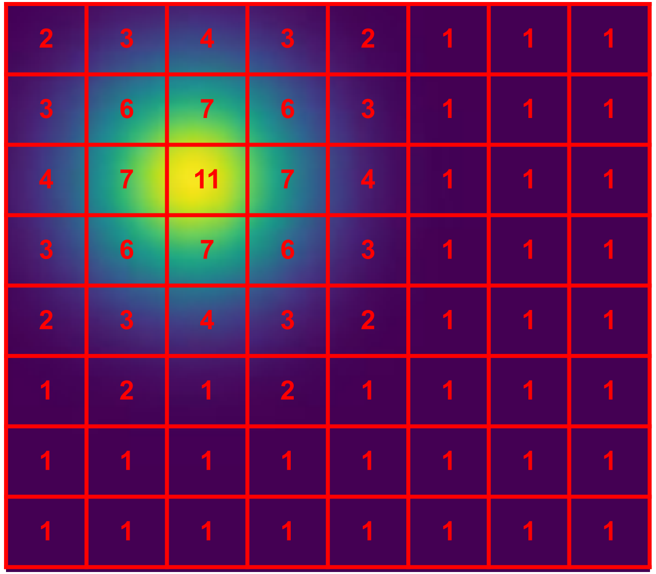
Method - Correspondence and Visibility Predictor

- $\varphi : \mathbb{R}^{H \times W \times 3} \times \mathbb{R}^{H \times W \times 3} \times \mathbb{R}^{H \times W} \to \mathbb{R}^{H \times W} \times \mathbb{R},$
  $\varphi(I_1, I_2, \mathcal{H}_{\mathrm{q}}) \to (\hat{\mathcal{H}}, \hat{v})$

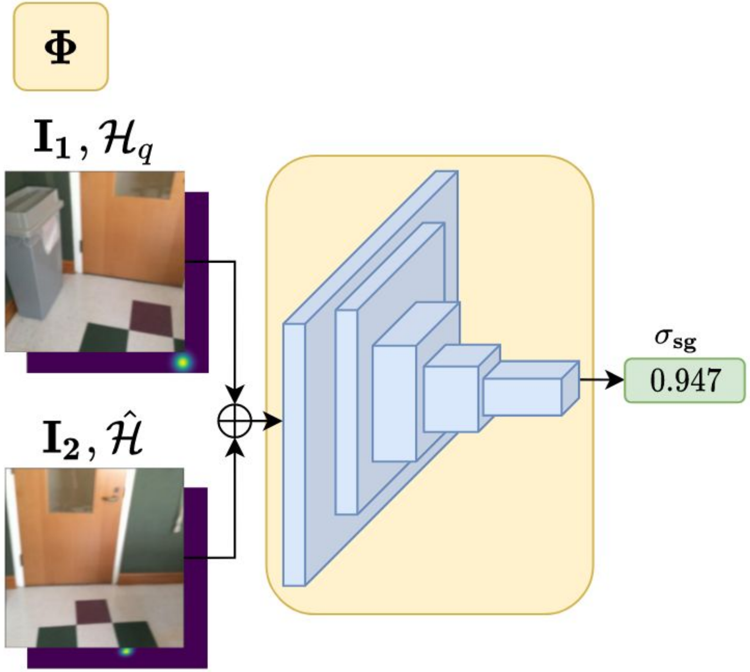
- $$\mathcal{L}_{\mathcal{H}} = \sum_i \Phi_{bce}(w_{\mathcal{H}}(\sigma_{\mathrm{sg}}(\mathcal{H}) - \mathcal{H}_{\mathrm{gt}})) + \lambda_{\mathrm{nll}} \sum_i \Phi_{nll}(w_{\mathcal{H}}(\sigma_{\mathrm{sm}}(\mathcal{H}) - \mathcal{H}_{\mathrm{gt}}))$$

- $w_{\mathcal{H}}(p) = 1 + 10G(m; \sigma = 7)(p)$

- $$\mathcal{L}_{\mathcal{V}} = \sum_i \Phi_{bce}\left(\hat{v} - v\right)$$



Method - Correspondence Weighting



Method - Differentiable Weighted Procrustes

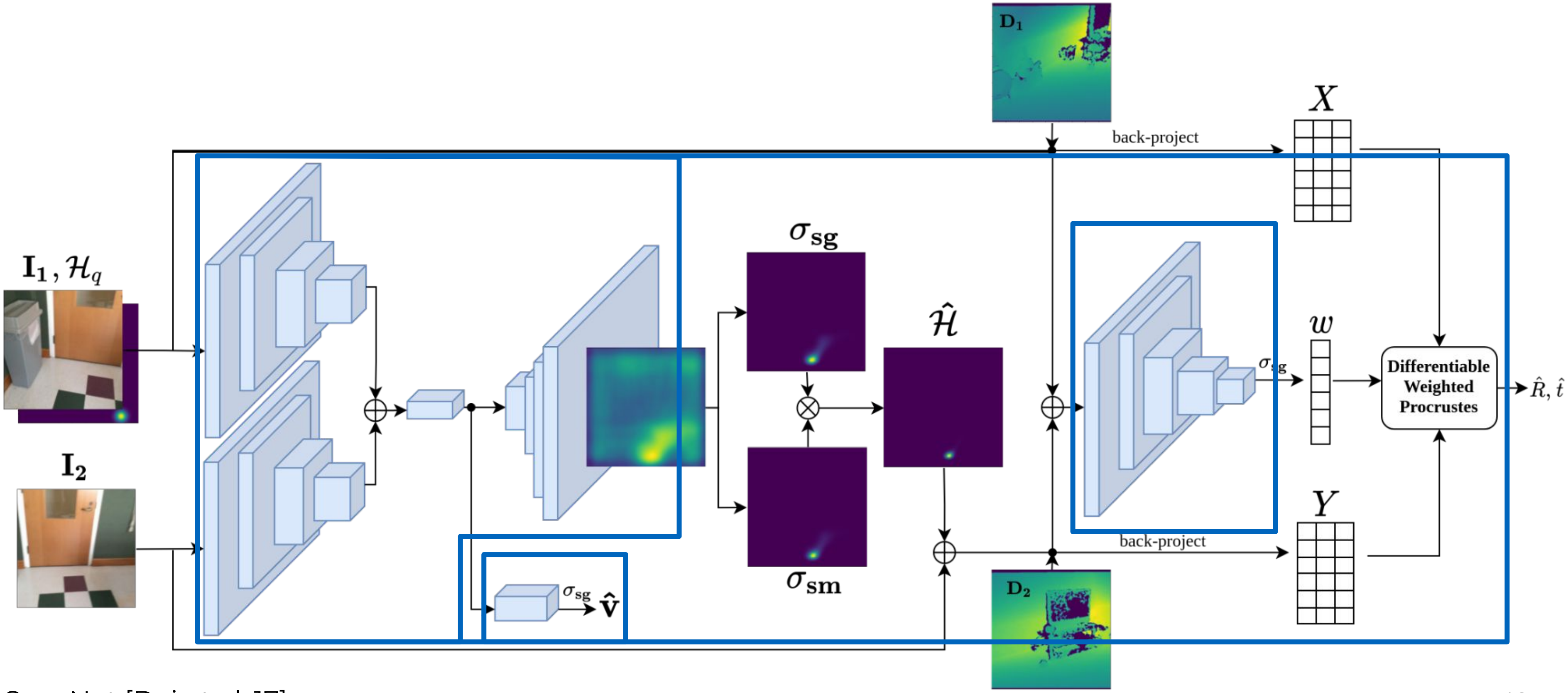
- X Back-projected visible query pixels
- Y Back-projected predicted matches
- W Predicted weights

- $\hat{R}, \hat{t} = argmin_{R,t} \sum w_i ||x_i - (Ry_i + t)||_2, R \in SO(3), t \in \mathbb{R}^3$

- $\mathcal{L}_{align} = ||R_p^T R_{gt} - I|| + ||t_p - t_{gt}||$

Deep Global Registration [Choy et al, 20], Deep Closest Point [Wang & Solomon, 19]

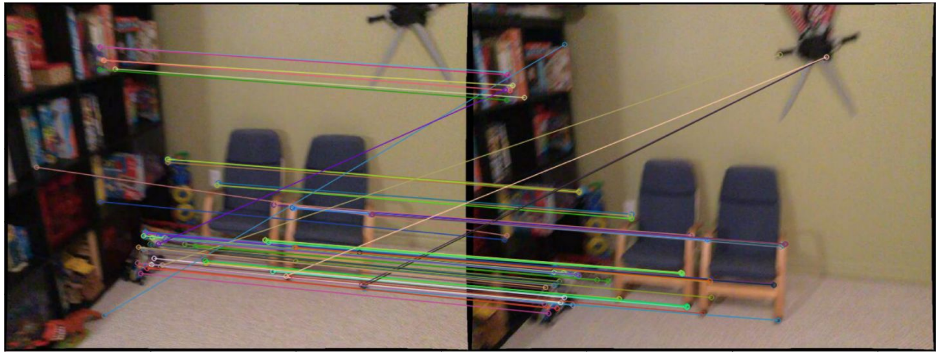
Training pipeline



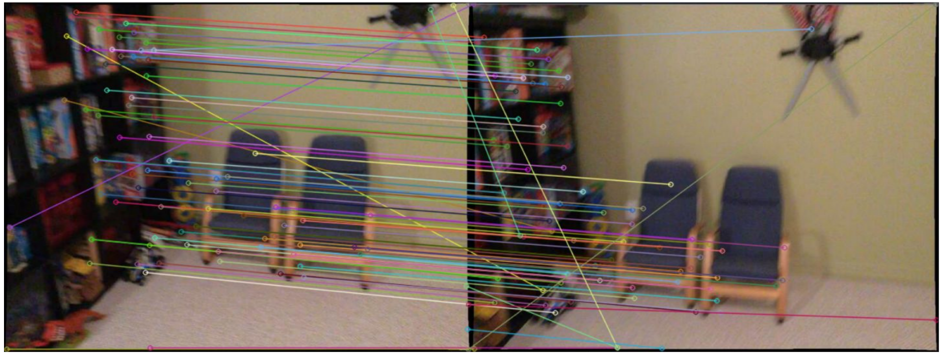
ScanNet [Dai et al, 17]

Results - Correspondence Prediction

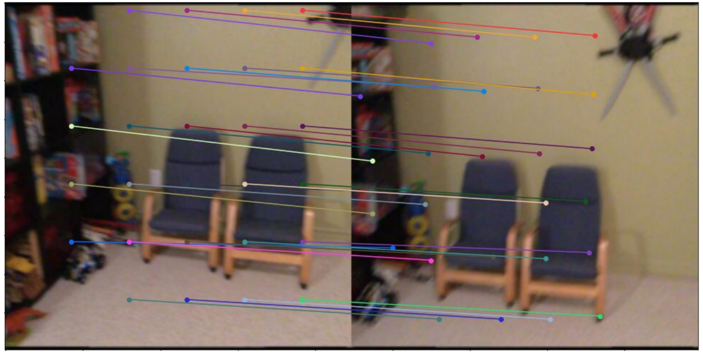
ORB



SIFT

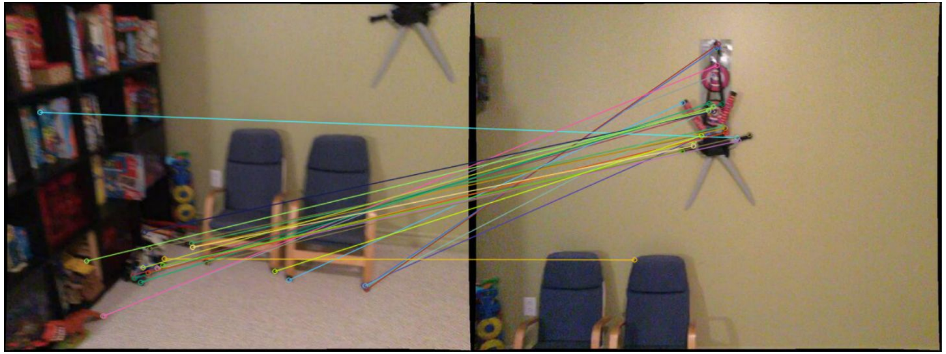


Model

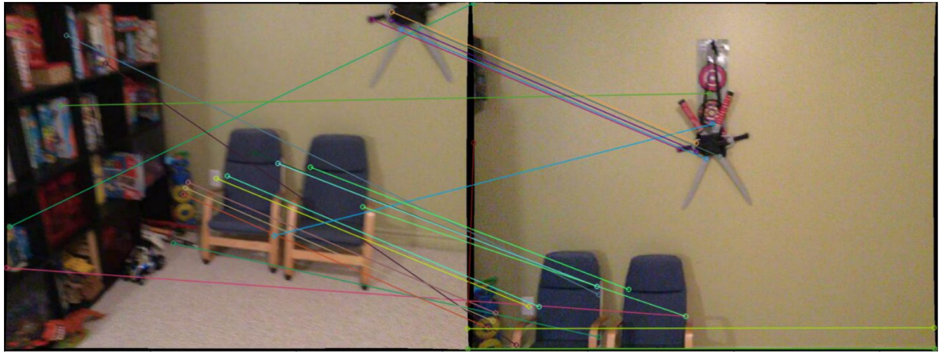


Results - Correspondence Prediction

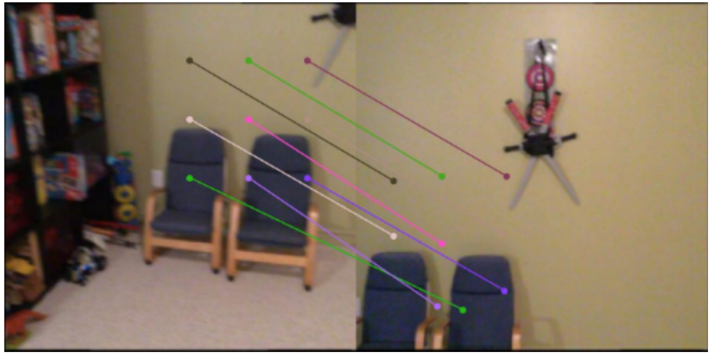
ORB



SIFT

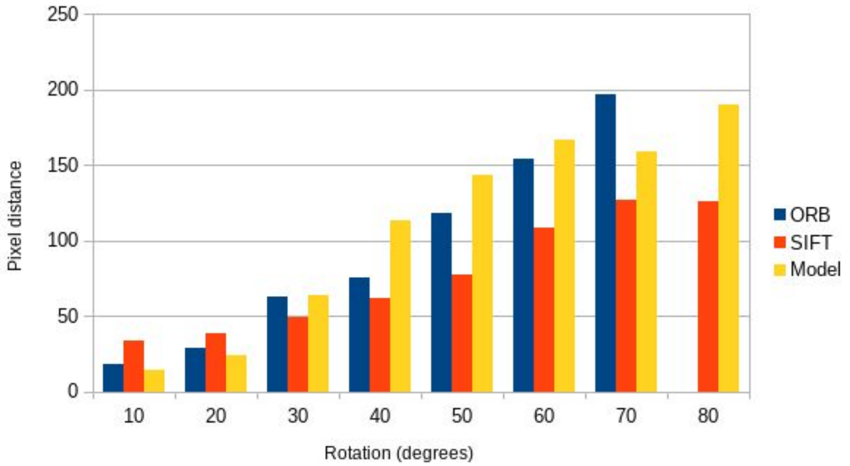


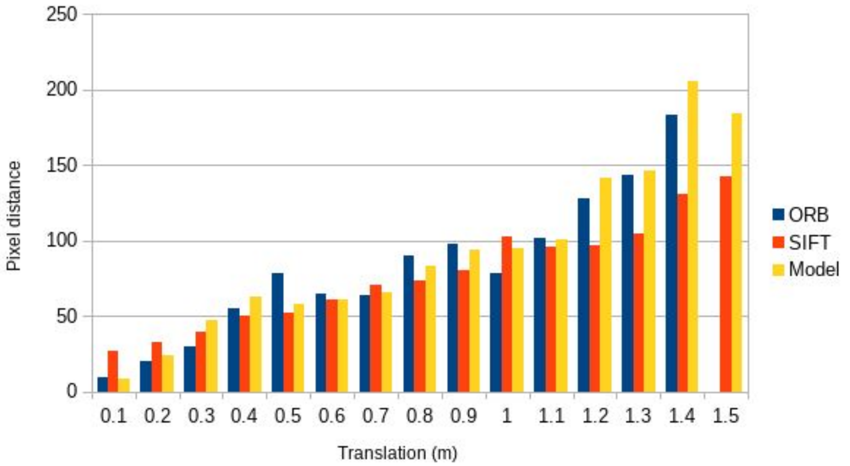
Model



Results - Correspondence Prediction

- Distance between ground truth and predicted correspondences

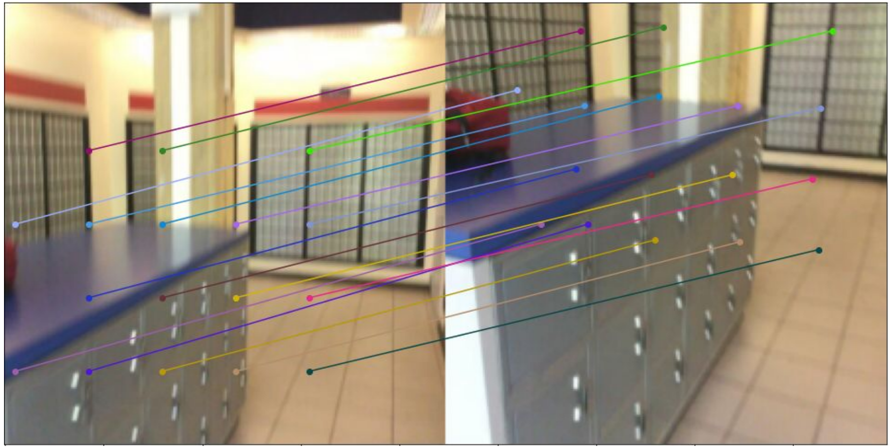


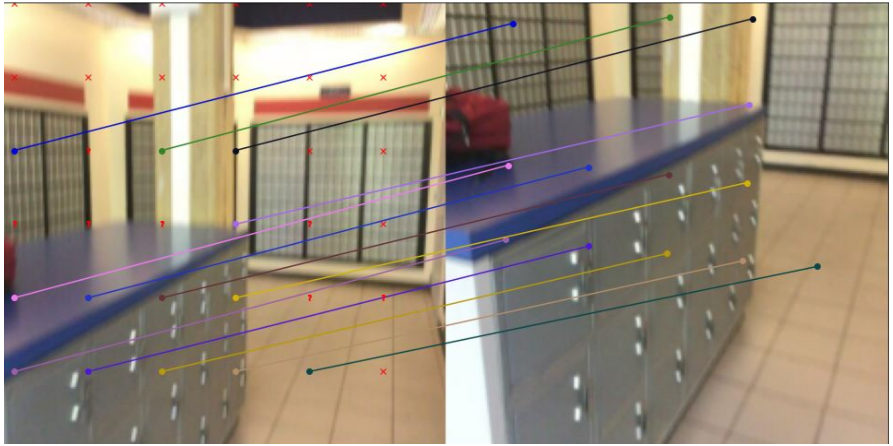


ORB[Rublee et al, 11], SIFT[Lindeberg et al, 12]

Results - Visibility Prediction

- Predicted matches
- Ground truth matches



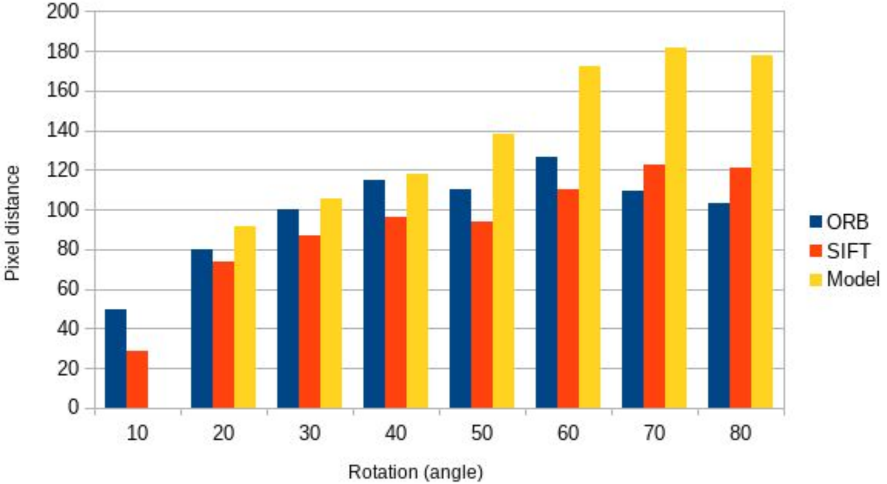


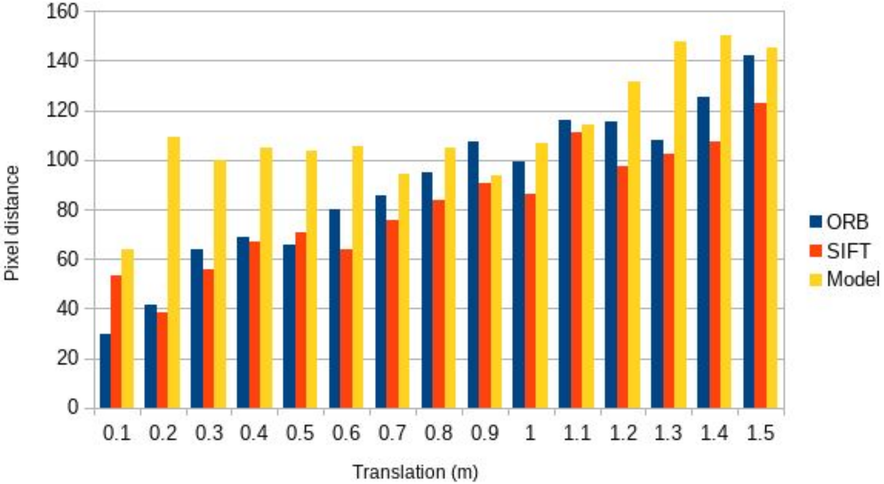
(x) - not visible, (?) - unknown

- Accuracy on validation dataset: 0.83

Results - Correspondence Prediction - 7 scenes

- Distance between ground truth and predicted correspondences

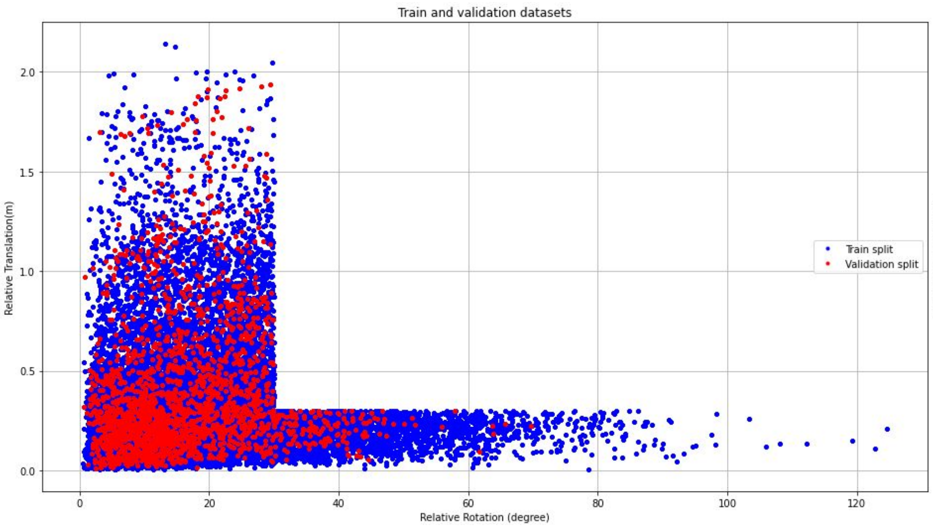




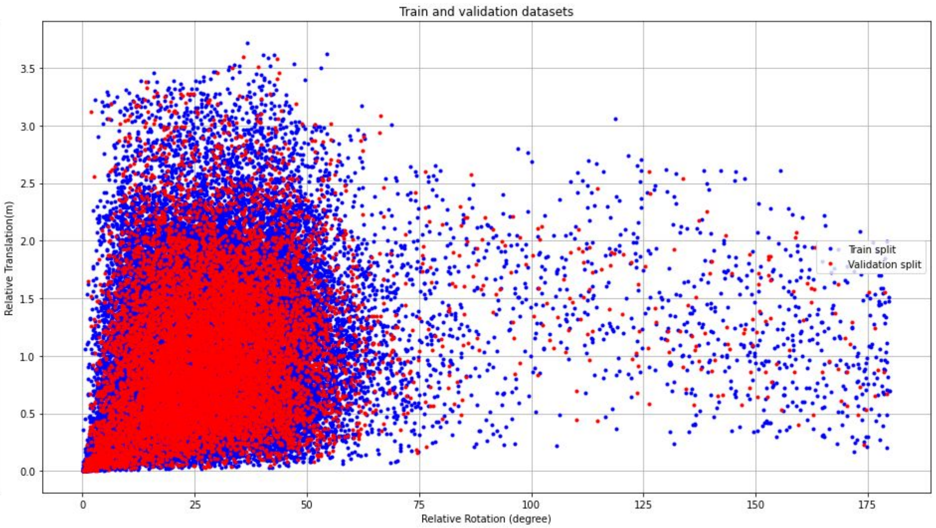
Real-time RGB-D camera relocalization[Glocker et al, 13], RelPoseNet[Melekhov et al, 17]

Relative Poses

- Our relative poses

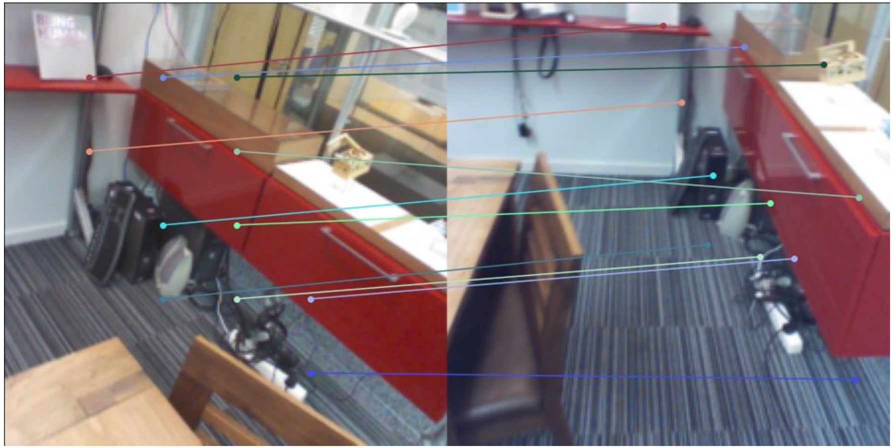


- Relative poses from 7-scenes dataset

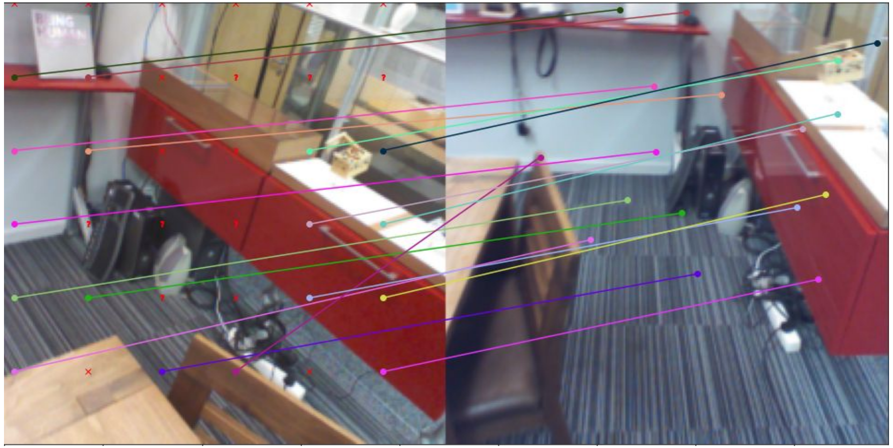


Results - Correspondence Prediction - 7 scenes

- Predicted matches



- Ground truth matches

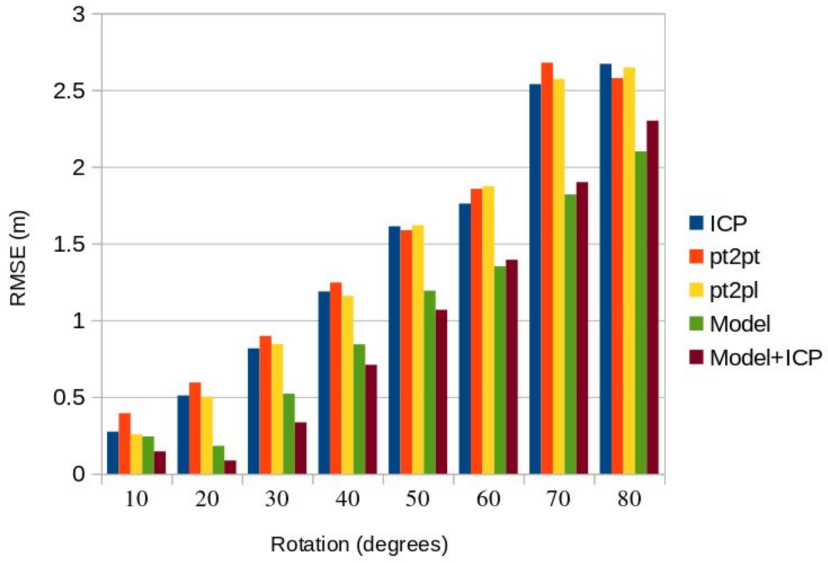


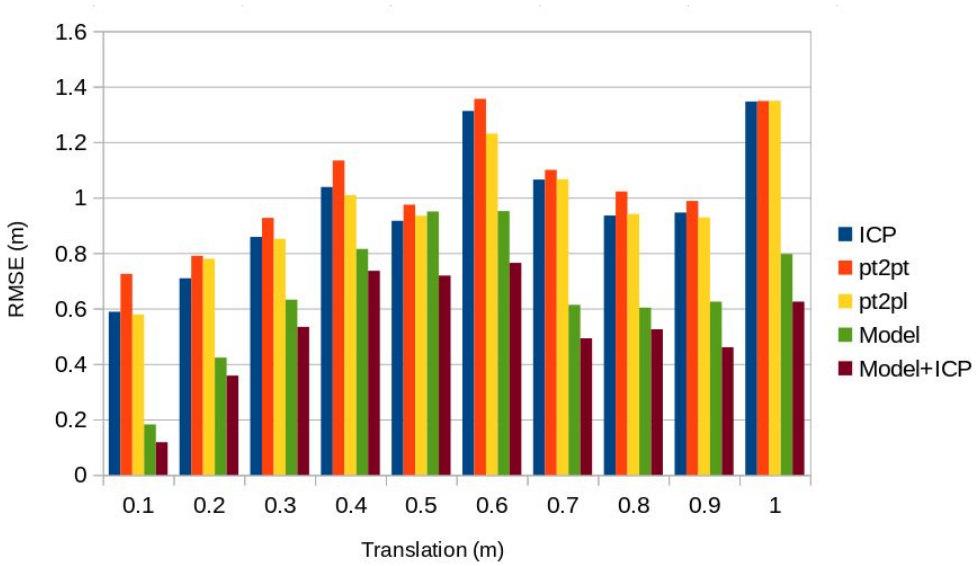
(**x**) - not visible

(**?**) - unknown

Results - 3D Reconstruction

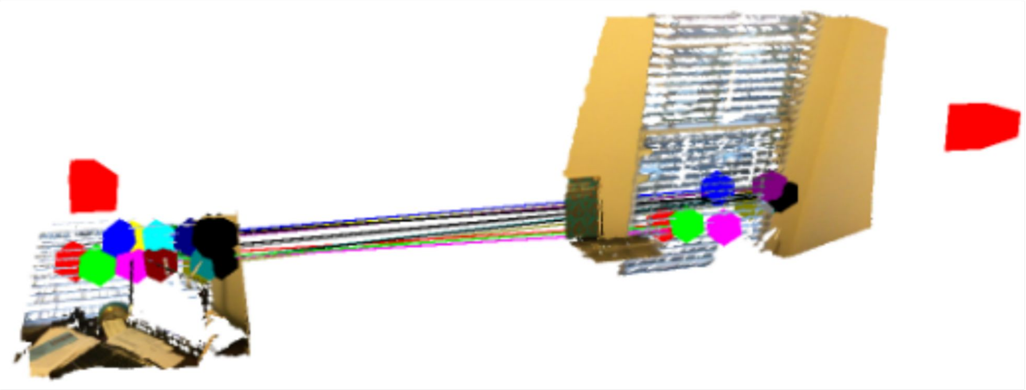
- Comparison of ICP, model and ICP with predicted pose as initial guess





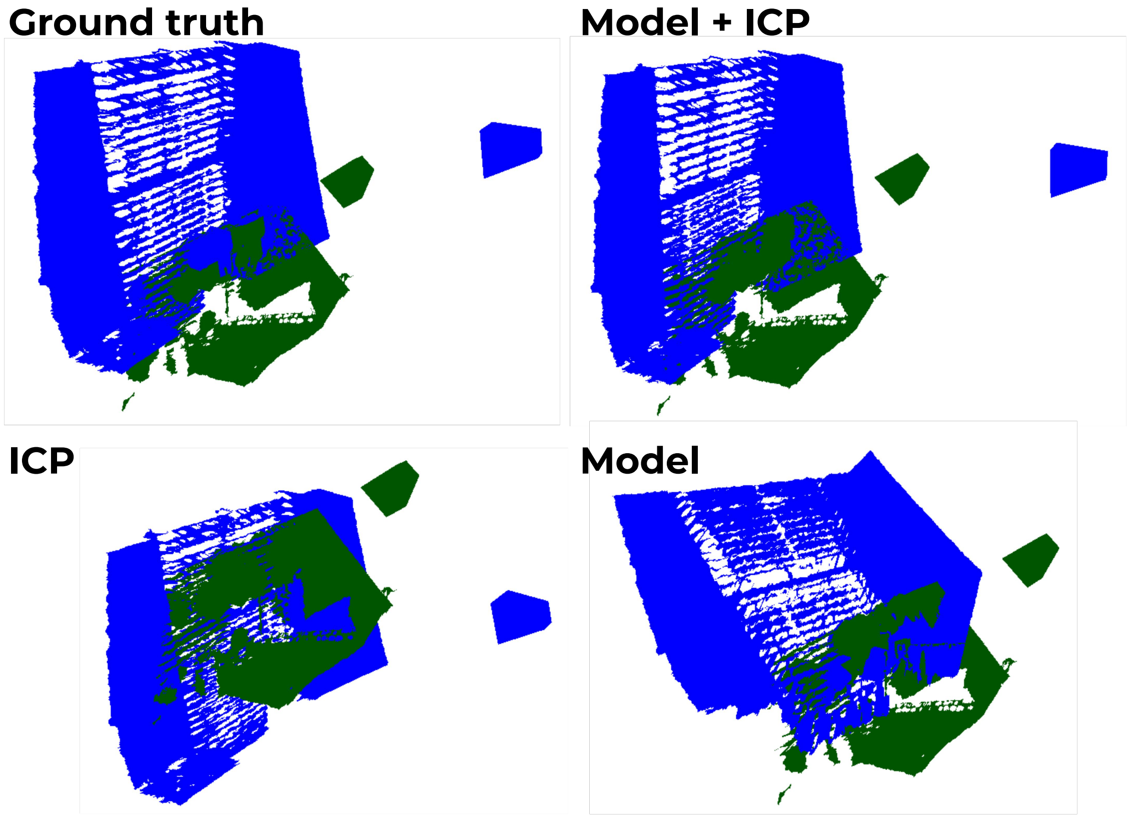
Results - 3D Reconstruction

- Target frame (left)
- Source frame (right)
- Predicted correspondences



Results - 3D Reconstruction

- Target frame (green)
- Source frame (blue)



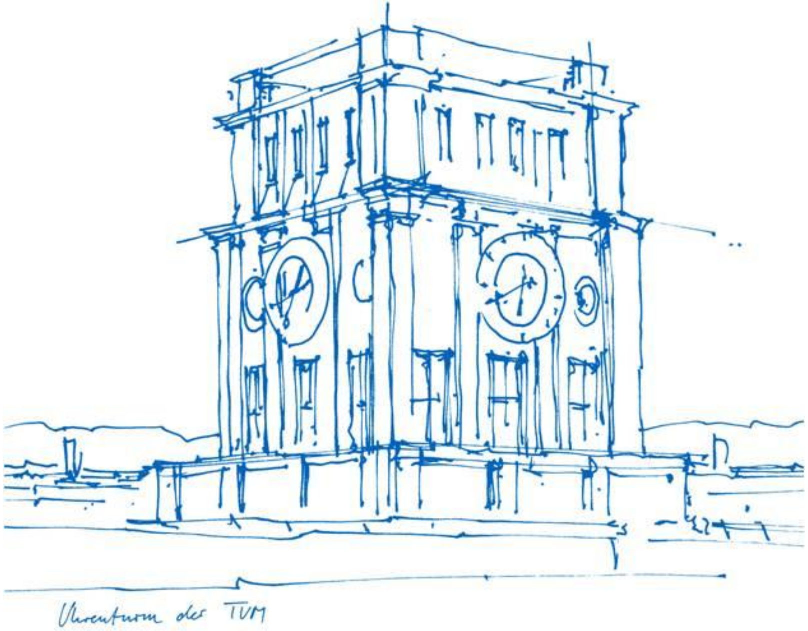
Conclusions

- We proposed a method for pairwise relative pose estimation
- It outperforms ICP for 3D scene alignment
- It helps avoiding local minima for ICP (better global solution when combined)

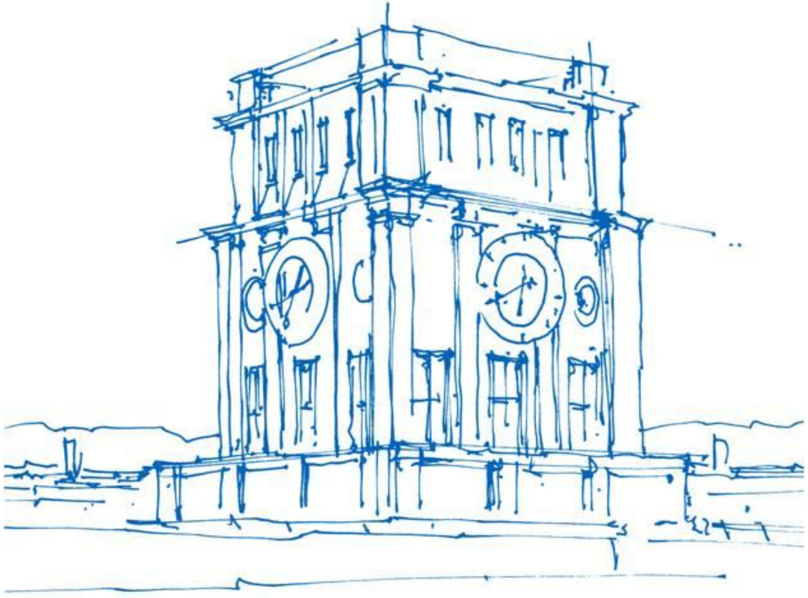
Future Work

- Correspondence weighting
- End to end
- Dense correspondences (for all pixels)

Questions



Appendix



Uhrenturm der TUM

Dataset

- We used ScanNet dataset
- For image id = i, pairs [(i, i + 20), (i, i + 50), (i, i + 80)] generated

Correspondence prediction

- Train
  - 380 scenes
  - 80k pairs of frames
  - 3.5M visibile matches
- Validation
  - 90 scenes
  - 20k pairs of frames
  - 800k visible matches

Visibility prediction

- Train
  - 90 scenes
  - 20k pairs of frames
  - 2M pairs of matches (800k visible, 1.2M occluded)
- Validation
  - 10 scenes
  - 2k pairs of frames
  - 220k matches (90k visible, 130k occluded)

Correspondence weighting

- Train
  - 1201 scenes
  - 307k pairs of frames
- Validation
  - 312 scenes
  - 80k pairs of frames

Training Details

	Correspondence prediction	Visibility prediction	Correspondence weighting*	End-to-end*
Learning rate	0.01	0.01	0.01	0.01
Momentum	0.9	0.9	0.9	0.9
Weight decay	1e-5	1e-5	1e-5	1e-5
Batch size	32	32	32	32
Iterations	60k	60k	-	-
Learning rate decay by 0.1	30k	53k	-	-
Training time	21h	14h	-	-

References

- [Besl & McKay ,92] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239-256, Feb. 1992, doi: 10.1109/34.121791.
- [Kendall et al, 15] Kendall, Alex, Matthew Grimes, and Roberto Cipolla. "Posenet: A convolutional network for real-time 6-dof camera relocalization." Proceedings of the IEEE international conference on computer vision. 2015.
- [Melekhov et al, 17] Melekhov, Iaroslav, et al. "Relative camera pose estimation using convolutional neural networks." International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, 2017.
- [Božič et al, 19] Bozic, Aljaz, et al. "Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- [Choy et al, 20] Choy, Christopher, Wei Dong, and Vladlen Koltun. "Deep global registration." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
- [Wang & Solomon, 19] Wang, Yue, and Justin M. Solomon. "Deep closest point: Learning representations for point cloud registration." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- [Dai et al, 17] Dai, Angela, et al. "Scannet: Richly-annotated 3d reconstructions of indoor scenes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- [Glocker et al, 13] Glocker, Ben, et al. "Real-time RGB-D camera relocalization." 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2013.