Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision





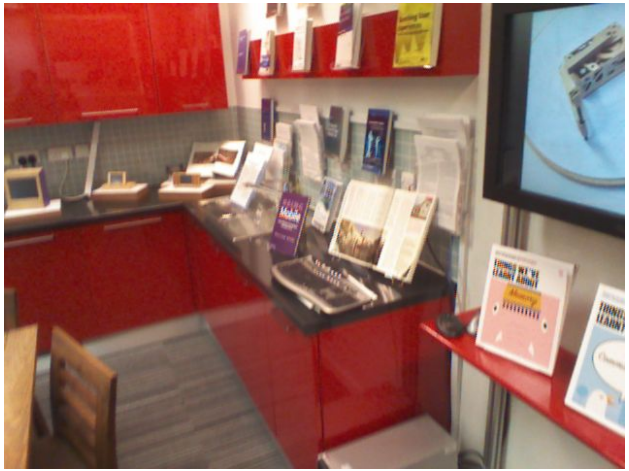
Mehmet Özgür Türkoğlu[1]* Eric Brachmann[2] Konrad Schindler[1] Gabriel J. Brostow[2,3] Áron Monszpart[2]

[1]ETH Zurich [2]Niantic [3]University College London

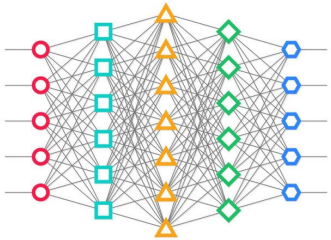
*Work done during an internship at Niantic.

ETH zürich NIANTIC UCL

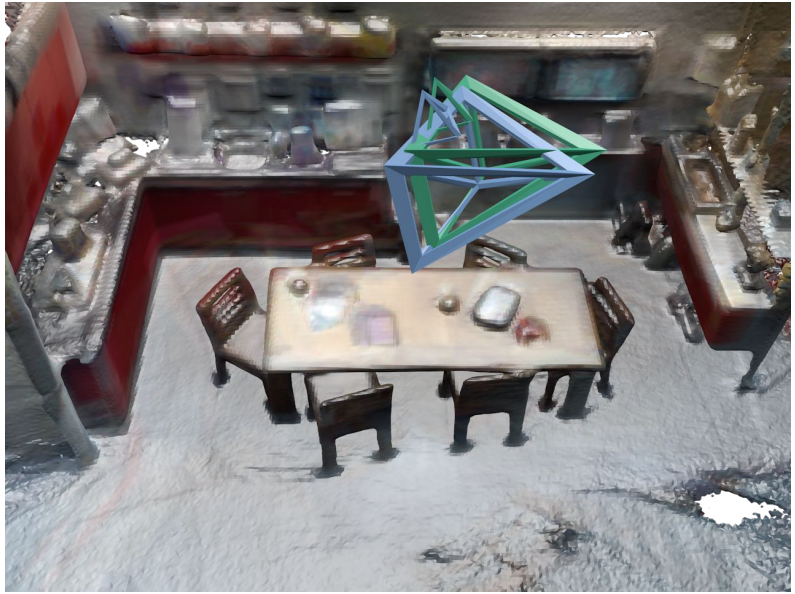
Visual Camera Re-localization



Query image



Algorithm e.g CNN-based pose regressor



6-DoF camera pose

Motivation

Structure-based methods [Sattler et al. 2019] achieve SOTA. So why look beyond structure-based methods?

- Camera intrinsics are often not available or reliable

[Sattler et al. 2019] Sattler et al. "Understanding the limitations of CNN-based absolute camera pose regression." In CVPR, 2019.

ETH zürich NIANTIC UCL

Motivation

Structure-based methods [Sattler et al. 2019] achieve SOTA. So why look beyond structure-based methods?

- Camera intrinsics are often not available or reliable
- Geometric optimization is costly

[Sattler et al. 2019] Sattler et al. "Understanding the limitations of CNN-based absolute camera pose regression." In CVPR, 2019.

ETH zürich NIANTIC UCL

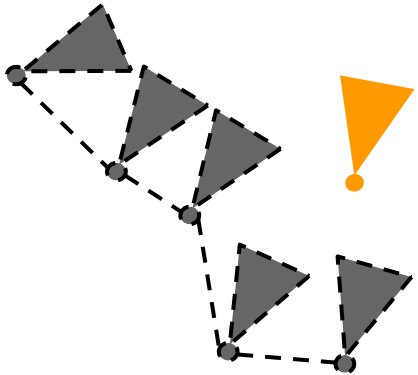
Motivation

Structure-based methods [Sattler et al. 2019] achieve SOTA. So why look beyond structure-based methods?

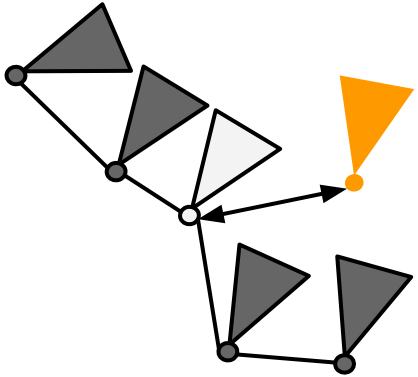
- Camera intrinsics are often not available or reliable
- Geometric optimization is costly
- Work best for scenes with easy-to track feature points

[Sattler et al. 2019] Sattler et al. "Understanding the limitations of CNN-based absolute camera pose regression." In CVPR, 2019.

Deep Absolute vs. Relative Pose Regression



APR *e.g.*, PoseNet



RPR *e.g.*, RelocNet

training pose

test pose

= only during training

= during both training/test

	APR	RPR
Scene-agnostic training	😩	💪
Generalize to unseen scene	😩	😎
Time complexity	💪	😐
Pose accuracy	😎	😐

😩=weak 😐=okay 😎=promising 💪= strong

[PoseNet] Kendall et al., "PoseNet: A convolutional network for real-time 6-dof camera relocalization." In ICCV 2015.
[RelocNet] Balntas et al., "RelocNet: Continuous metric learning relocalisation using neural nets." In ECCV 2018.

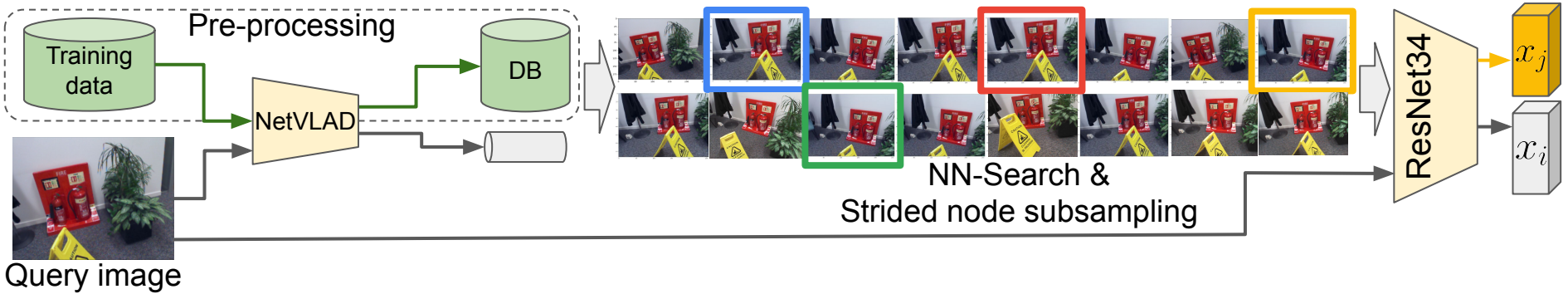
ETH zürich NIANTIC UCL

Method Overview

- Image retrieval + GNN
- Supervised with only relative poses of training scenes

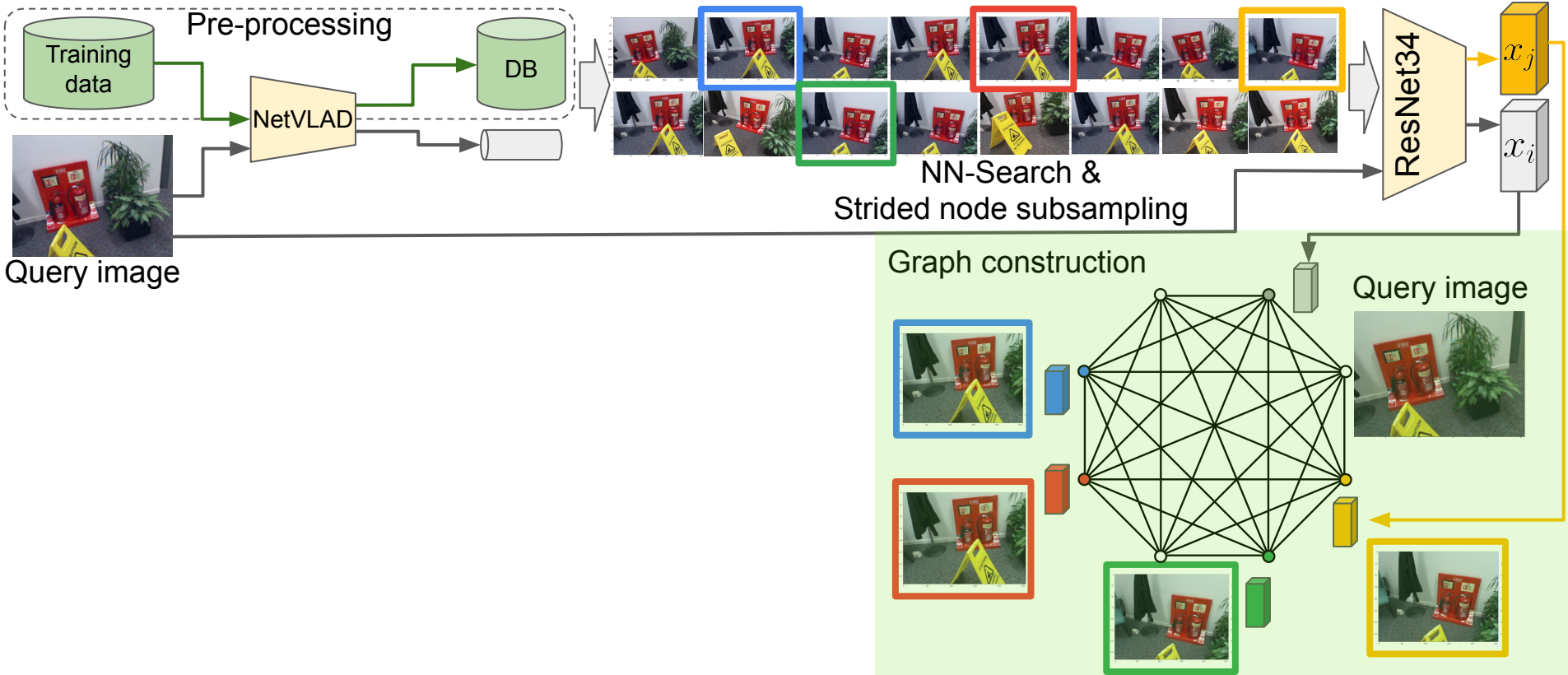
Method Overview

- Image retrieval + GNN
- Supervised with only relative poses of training scenes



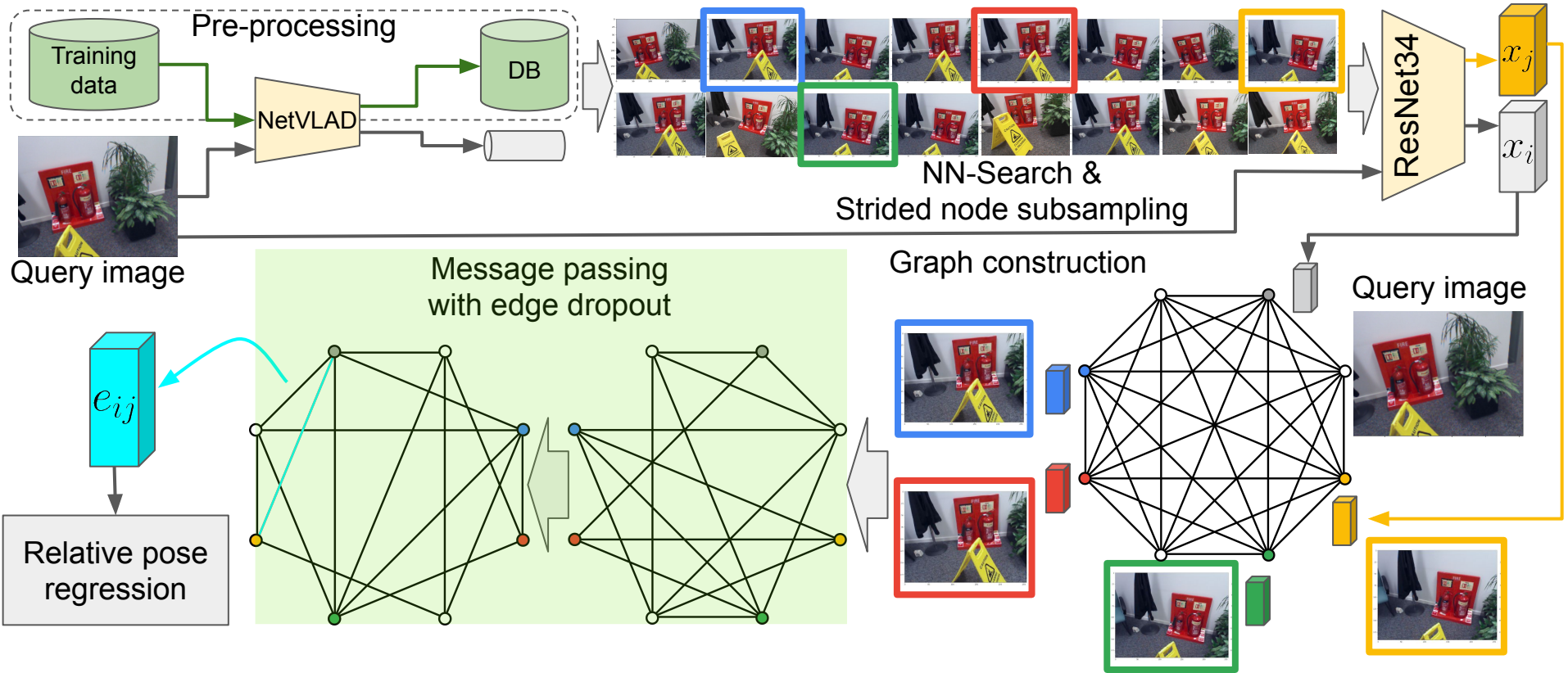
Method Overview

- Image retrieval + GNN
- Supervised with only relative poses of training scenes



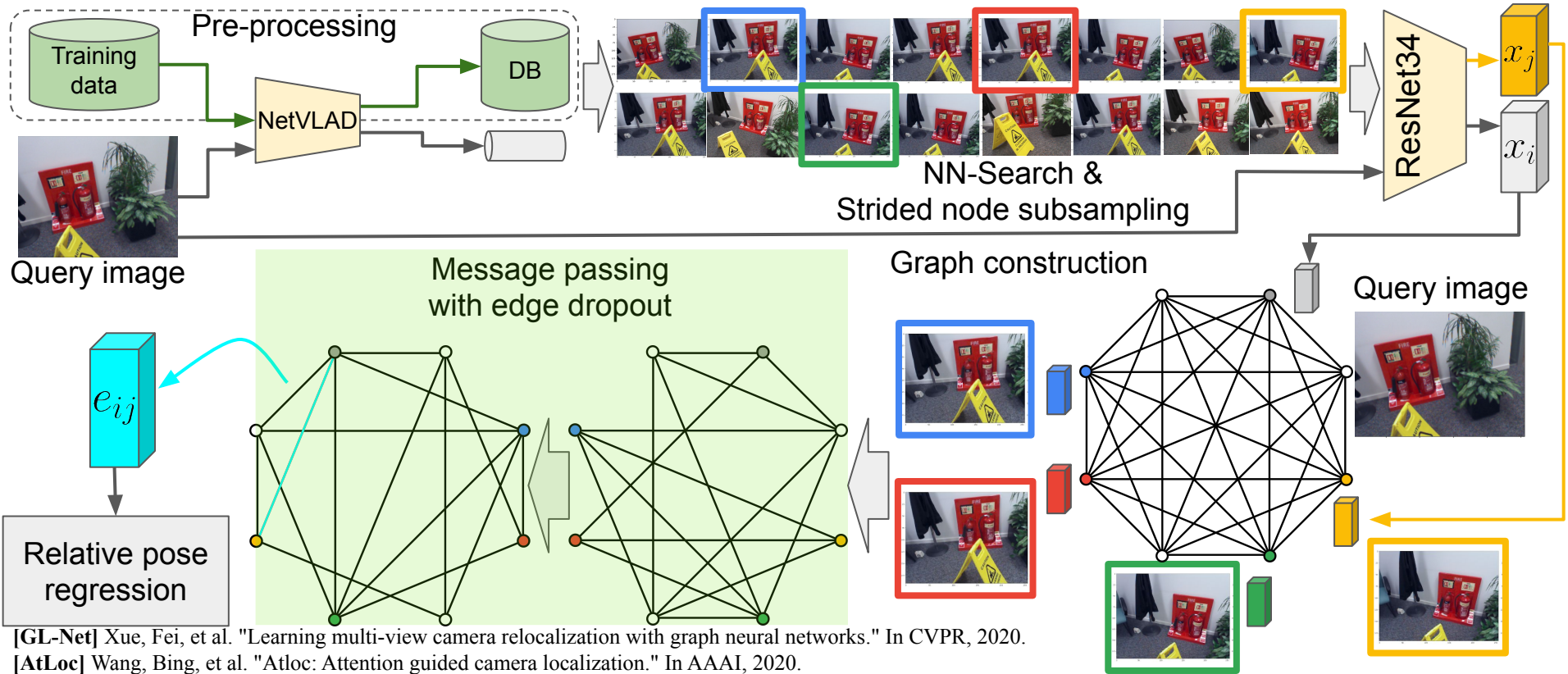
Method Overview

- Image retrieval + GNN
- Supervised with only relative poses of training scenes



Method Overview

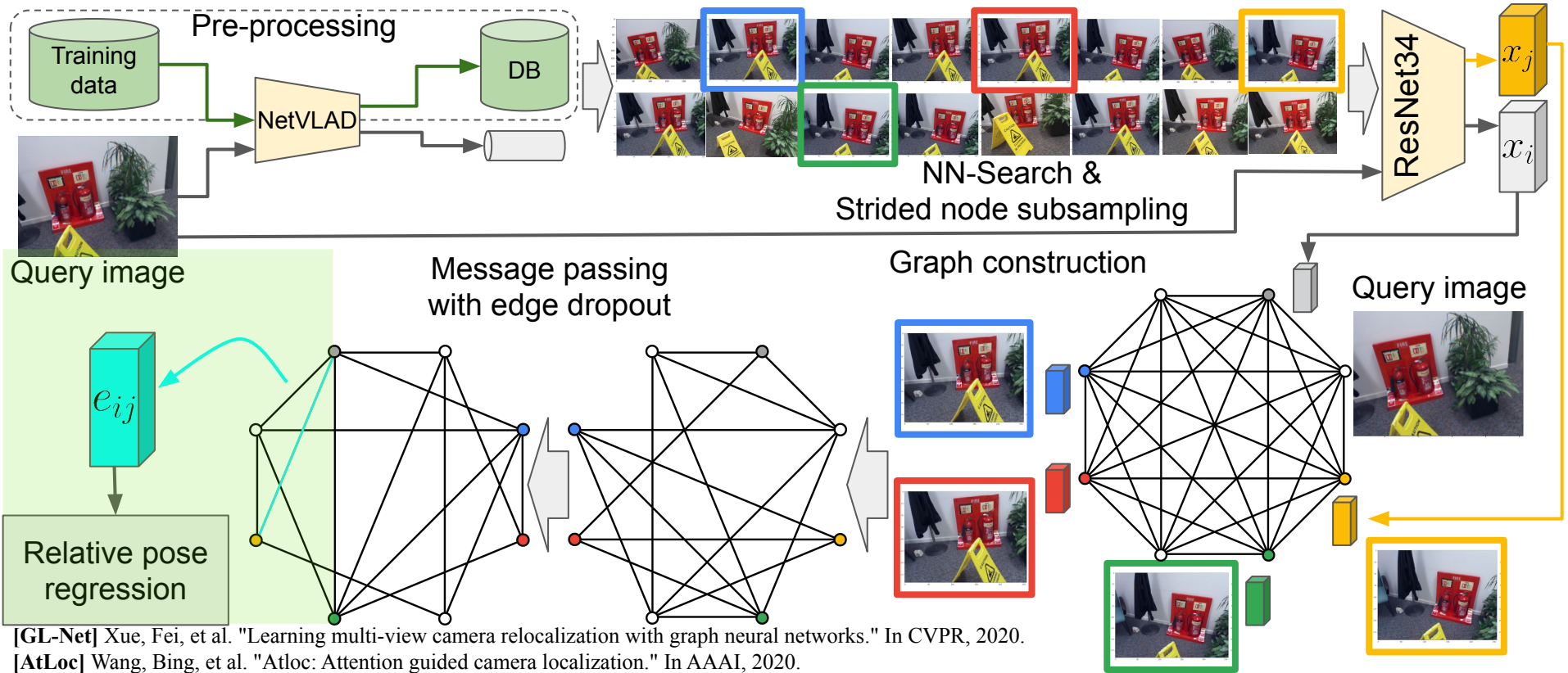
- Image retrieval + GNN
- Supervised with only relative poses of training scenes



[GL-Net] Xue, Fei, et al. "Learning multi-view camera relocalization with graph neural networks." In CVPR, 2020.
[AtLoc] Wang, Bing, et al. "Atloc: Attention guided camera localization." In AAAI, 2020.

Method Overview

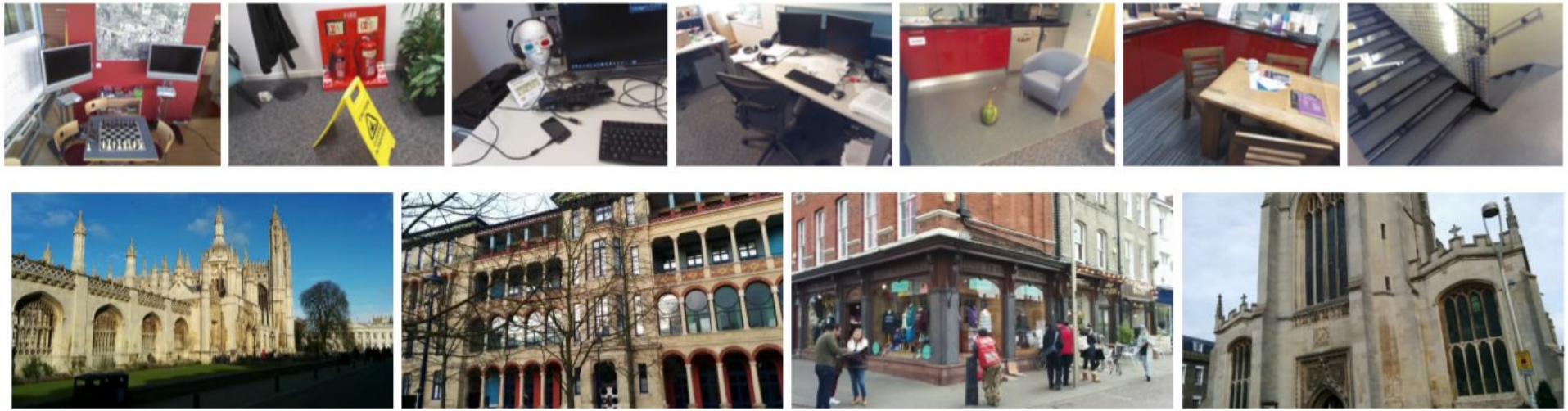
- Image retrieval + GNN
- Supervised with only relative poses of training scenes



[GL-Net] Xue, Fei, et al. "Learning multi-view camera relocalization with graph neural networks." In CVPR, 2020.
[AtLoc] Wang, Bing, et al. "Atloc: Attention guided camera localization." In AAAI, 2020.

Experiments

- Datasets
  - Indoor: 7-Scenes
  - Outdoor: Cambridge Landmarks
- Evaluation metric
  - Median translation and rotation error



Results: 7-Scenes

		# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
	DSAC* [11]*	1	0.02, 1.1°	0.02, 1.2°	0.01, 1.8°	0.03, 1.2°	0.04, 1.4°	0.03, 1.7°	0.04, 1.4°	0.03, 1.4°
Seq. based	VidLoc [21]*	200	0.18, -	0.26, -	0.14, -	0.26, -	0.36, -	0.31, -	0.26, -	0.25, -
	LsG [69]*	7	0.09, 3.3°	0.26, 10.9°	0.17, 12.7°	0.18, 5.5°	0.20, 3.7°	0.23, 4.9°	0.23, 11.3°	0.19, 7.5°
	MapNet [12]*	3	0.08, 3.3°	0.27, 11.7°	0.18, 13.3°	0.17, 5.2°	0.22, 4.0°	0.23, 4.9°	0.30, 12.1°	0.21, 7.8°
	GL-Net [70]*?	8	0.08, 2.8°	0.26, 8.9°	0.17, 11.4°	0.18, 5.1°	0.15, 2.8°	0.25, 4.5°	0.23, 8.8°	0.19, 6.3°
Image based APR	PoseNet [36]*	1	0.32, 6.6°	0.47, 14.0°	0.30, 12.2°	0.48, 7.2°	0.49, 8.1°	0.58, 8.3°	0.48, 13.1°	0.45, 9.9°
	Bayesian PoseNet [34]*	1	0.37, 7.2°	0.43, 13.7°	0.31, 12.0°	0.48, 8.0°	0.61, 7.1°	0.58, 7.5°	0.48, 13.1°	0.47, 9.8°
	Geometric PoseNet [35]*	1	0.13, 4.5°	0.27, 11.3°	0.17, 13.0°	0.19, 5.6°	0.26, 4.8°	0.23, 5.4°	0.35, 12.4°	0.23, 8.1°
	MLFBPPose [66]*	1	0.12, 5.8°	0.26, 12.0°	0.14, 13.5°	0.18, 8.2°	0.21, 7.1°	0.22, 8.1°	0.26, 13.6°	0.20, 9.8°
	Hourglass [44]*	1	0.15, 6.2°	0.27, 10.8°	0.19, 11.6°	0.21, 8.5°	0.25, 7.0°	0.27, 10.2°	0.29, 12.5°	0.23, 9.5°
	LSTM-Pose [63]*	1	0.24, 5.8°	0.34, 11.9°	0.21, 13.7°	0.30, 8.1°	0.33, 7.0°	0.37, 8.8°	0.40, 13.7°	0.31, 9.9°
	BranchNet [67]*	1	0.18, 5.2°	0.34, 9.0°	0.20, 14.2°	0.30, 7.1°	0.27, 5.1°	0.33, 7.4°	0.38, 10.3°	0.29, 8.3°
	ANNet [13]*	1	0.12, 4.3°	0.27, 11.6°	0.16, 12.4°	0.19, 6.8°	0.21, 5.2°	0.25, 6.0°	0.28, 8.4°	0.21, 7.9°
	GPoseNet [14]*	1	0.20, 7.1°	0.38, 12.3°	0.21, 13.8°	0.28, 8.8°	0.37, 6.9°	0.35, 8.2°	0.37, 12.5°	0.31, 10.0°
	AttLoc [64]*	1	0.10, 4.1°	0.25, 11.4°	0.16, 11.8°	0.17, 5.3°	0.21, 4.4°	0.23, 5.4°	0.26, 10.5°	0.20, 7.6°
	AnchorPoint [48]*?	1	**0.06**, 3.9°	**0.16**, 11.1°	**0.09**, 11.2°	**0.11**, 5.4°	**0.14**, 3.6°	**0.13**, 5.3°	**0.21**, 11.9°	0.13, 7.5°
IR	DenseVLAD [58]	1	0.21, 12.5°	0.33, 13.8°	0.15, 14.9°	0.28, 11.2°	0.31, 11.2°	0.30, 11.3°	0.25, 12.3°	0.26, 12.5°
	DenseVLAD+Inter [54]	1	0.18, 10.0°	0.33, 12.4°	0.14, 14.3°	0.25, 10.1°	0.26, 9.4°	0.27, 11.1°	0.24, 14.7°	0.24, 11.7°
RPR	NN-Net [38]	1	0.13, 6.5°	0.26, 12.7°	0.14, 12.3°	0.21, 7.4°	0.24, 6.4°	0.24, 8.0°	0.27, 11.8°	0.21, 9.3°
	RelocNet [3]	1	0.12, 4.1°	0.26, 10.4°	0.14, 10.5°	0.18, 5.3°	0.26, 4.2°	0.23, 5.1°	0.28, 7.5°	0.21, 6.7°
	EssNet [72]	1	0.13, 5.1°	0.27, 10.1°	0.15, 9.9°	0.21, 6.9°	0.22, 6.1°	0.23, 6.9°	0.32, 11.2°	0.22, 8.0°
	EssNet [72] reprod.	1	-	-	-	-	-	-	0.32, 9.8°	-
	NC-EssNet [72]	1	0.12, 5.6°	0.26, 9.6°	0.14, 10.7°	0.20 , 6.7°	0.22, 5.7°	0.22, 6.3°	0.31, 7.9°	0.21, 7.5°
	NC-EssNet [72] reprod.	1	0.13, 5.5°	-	-	-	-	-	-	-
	CamNet [24]?	1	-	-	-	-	-	-	-	**0.05**, **1.8°**
	Ours	1	0.09, **2.7°**	0.24, **7.5°**	0.13, **8.7°**	0.15, **4.1°**	0.17, **3.5°**	0.20, **3.7°**	0.23, **6.5°**	0.17, 5.2°

= main competitor/ours = baseline unreproducible w/public version of the code

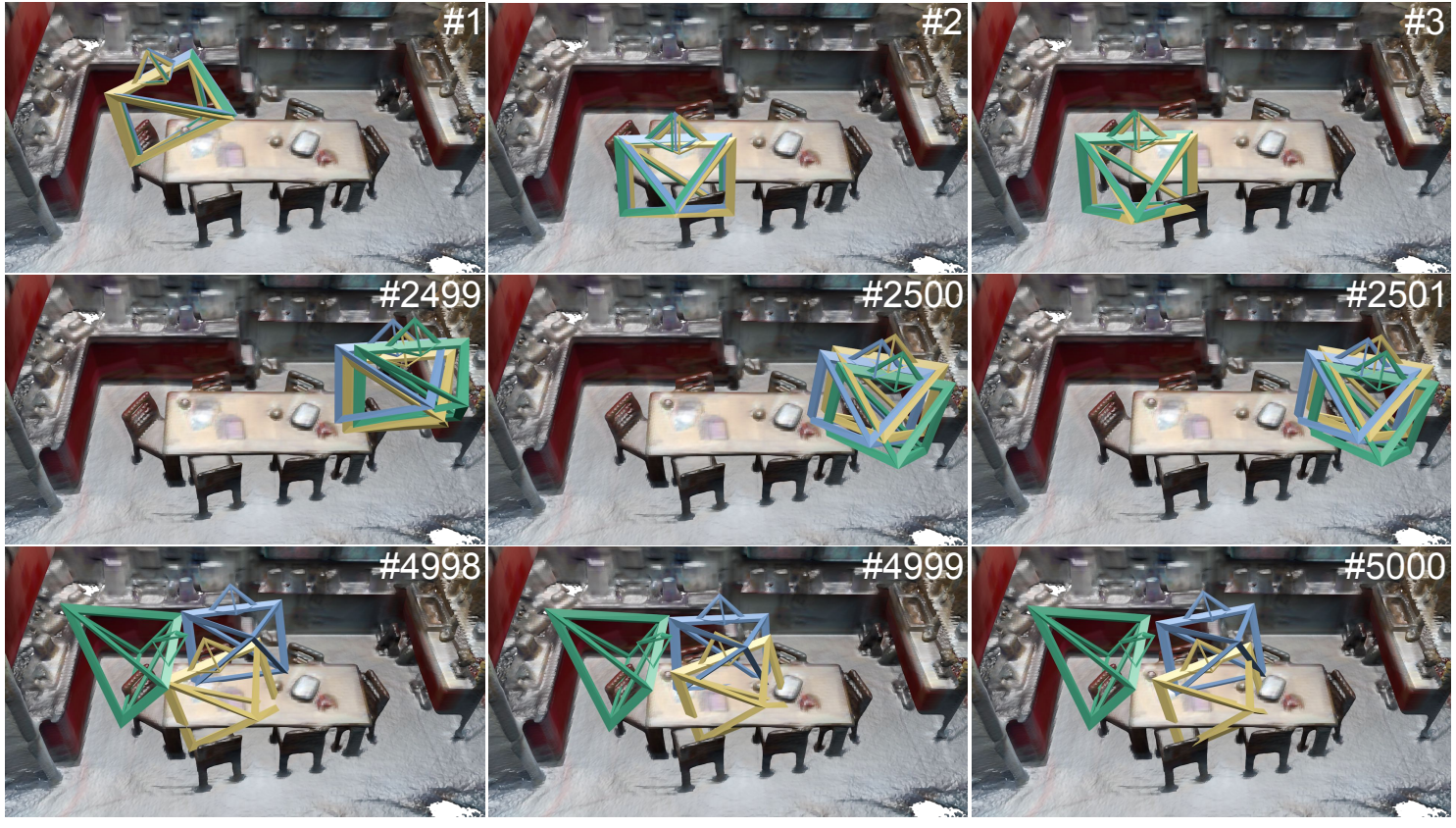
Results: 7-Scenes

		# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
	DSAC* [11]*	1	0.02, 1.1°	0.02, 1.2°	0.01, 1.8°	0.03, 1.2°	0.04, 1.4°	0.03, 1.7°	0.04, 1.4°	0.03, 1.4°
Seq. based	VidLoc [21]*	200	0.18, -	0.26, -	0.14, -	0.26, -	0.36, -	0.31, -	0.26, -	0.25, -
	LsG [69]*	7	0.09, 3.3°	0.26, 10.9°	0.17, 12.7°	0.18, 5.5°	0.20, 3.7°	0.23, 4.9°	0.23, 11.3°	0.19, 7.5°
	MapNet [12]*	3	0.08, 3.3°	0.27, 11.7°	0.18, 13.3°	0.17, 5.2°	0.22, 4.0°	0.23, 4.9°	0.30, 12.1°	0.21, 7.8°
	GL-Net [70]*?	8	0.08, 2.8°	0.26, 8.9°	0.17, 11.4°	0.18, 5.1°	0.15, 2.8°	0.25, 4.5°	0.23, 8.8°	0.19, 6.3°
Image based APR	PoseNet [36]*	1	0.32, 6.6°	0.47, 14.0°	0.30, 12.2°	0.48, 7.2°	0.49, 8.1°	0.58, 8.3°	0.48, 13.1°	0.45, 9.9°
	Bayesian PoseNet [34]*	1	0.37, 7.2°	0.43, 13.7°	0.31, 12.0°	0.48, 8.0°	0.61, 7.1°	0.58, 7.5°	0.48, 13.1°	0.47, 9.8°
	Geometric PoseNet [35]*	1	0.13, 4.5°	0.27, 11.3°	0.17, 13.0°	0.19, 5.6°	0.26, 4.8°	0.23, 5.4°	0.35, 12.4°	0.23, 8.1°
	MLFBPPose [66]*	1	0.12, 5.8°	0.26, 12.0°	0.14, 13.5°	0.18, 8.2°	0.21, 7.1°	0.22, 8.1°	0.26, 13.6°	0.20, 9.8°
	Hourglass [44]*	1	0.15, 6.2°	0.27, 10.8°	0.19, 11.6°	0.21, 8.5°	0.25, 7.0°	0.27, 10.2°	0.29, 12.5°	0.23, 9.5°
	LSTM-Pose [63]*	1	0.24, 5.8°	0.34, 11.9°	0.21, 13.7°	0.30, 8.1°	0.33, 7.0°	0.37, 8.8°	0.40, 13.7°	0.31, 9.9°
	BranchNet [67]*	1	0.18, 5.2°	0.34, 9.0°	0.20, 14.2°	0.30, 7.1°	0.27, 5.1°	0.33, 7.4°	0.38, 10.3°	0.29, 8.3°
	ANNet [13]*	1	0.12, 4.3°	0.27, 11.6°	0.16, 12.4°	0.19, 6.8°	0.21, 5.2°	0.25, 6.0°	0.28, 8.4°	0.21, 7.9°
	GPoseNet [14]*	1	0.20, 7.1°	0.38, 12.3°	0.21, 13.8°	0.28, 8.8°	0.37, 6.9°	0.35, 8.2°	0.37, 12.5°	0.31, 10.0°
	AttLoc [64]*	1	0.10, 4.1°	0.25, 11.4°	0.16, 11.8°	0.17, 5.3°	0.21, 4.4°	0.23, 5.4°	0.26, 10.5°	0.20, 7.6°
	AnchorPoint [48]*?	1	**0.06**, 3.9°	**0.16**, 11.1°	**0.09**, 11.2°	**0.11**, 5.4°	**0.14**, 3.6°	**0.13**, 5.3°	**0.21**, 11.9°	0.13, 7.5°
IR	DenseVLAD [58]	1	0.21, 12.5°	0.33, 13.8°	0.15, 14.9°	0.28, 11.2°	0.31, 11.2°	0.30, 11.3°	0.25, 12.3°	0.26, 12.5°
	DenseVLAD+Inter [54]	1	0.18, 10.0°	0.33, 12.4°	0.14, 14.3°	0.25, 10.1°	0.26, 9.4°	0.27, 11.1°	0.24, 14.7°	0.24, 11.7°
RPR	NN-Net [38]	1	0.13, 6.5°	0.26, 12.7°	0.14, 12.3°	0.21, 7.4°	0.24, 6.4°	0.24, 8.0°	0.27, 11.8°	0.21, 9.3°
	RelocNet [3]	1	0.12, 4.1°	0.26, 10.4°	0.14, 10.5°	0.18, 5.3°	0.26, 4.2°	0.23, 5.1°	0.28, 7.5°	0.21, 6.7°
	EssNet [72]	1	0.13, 5.1°	0.27, 10.1°	0.15, 9.9°	0.21, 6.9°	0.22, 6.1°	0.23, 6.9°	0.32, 11.2°	0.22, 8.0°
	EssNet [72] reprod.	1	-	-	-	-	-	-	0.32, 9.8°	-
	NC-EssNet [72]	1	0.12, 5.6°	0.26, 9.6°	0.14, 10.7°	0.20 , 6.7°	0.22, 5.7°	0.22, 6.3°	0.31, 7.9°	0.21, 7.5°
	NC-EssNet [72] reprod.	1	0.13, 5.5°	-	-	-	-	-	-	-
	CamNet [24]?	1	-	-	-	-	-	-	-	**0.05**, **1.8°**
	Ours	1	0.08, **2.7°**	0.21, **7.5°**	0.13, **8.7°**	0.15, **4.1°**	0.15, **3.5°**	0.19, **3.7°**	0.22, **6.5°**	0.16, 5.2°

= main competitor/ours = unreproducible baseline w/public version of the code

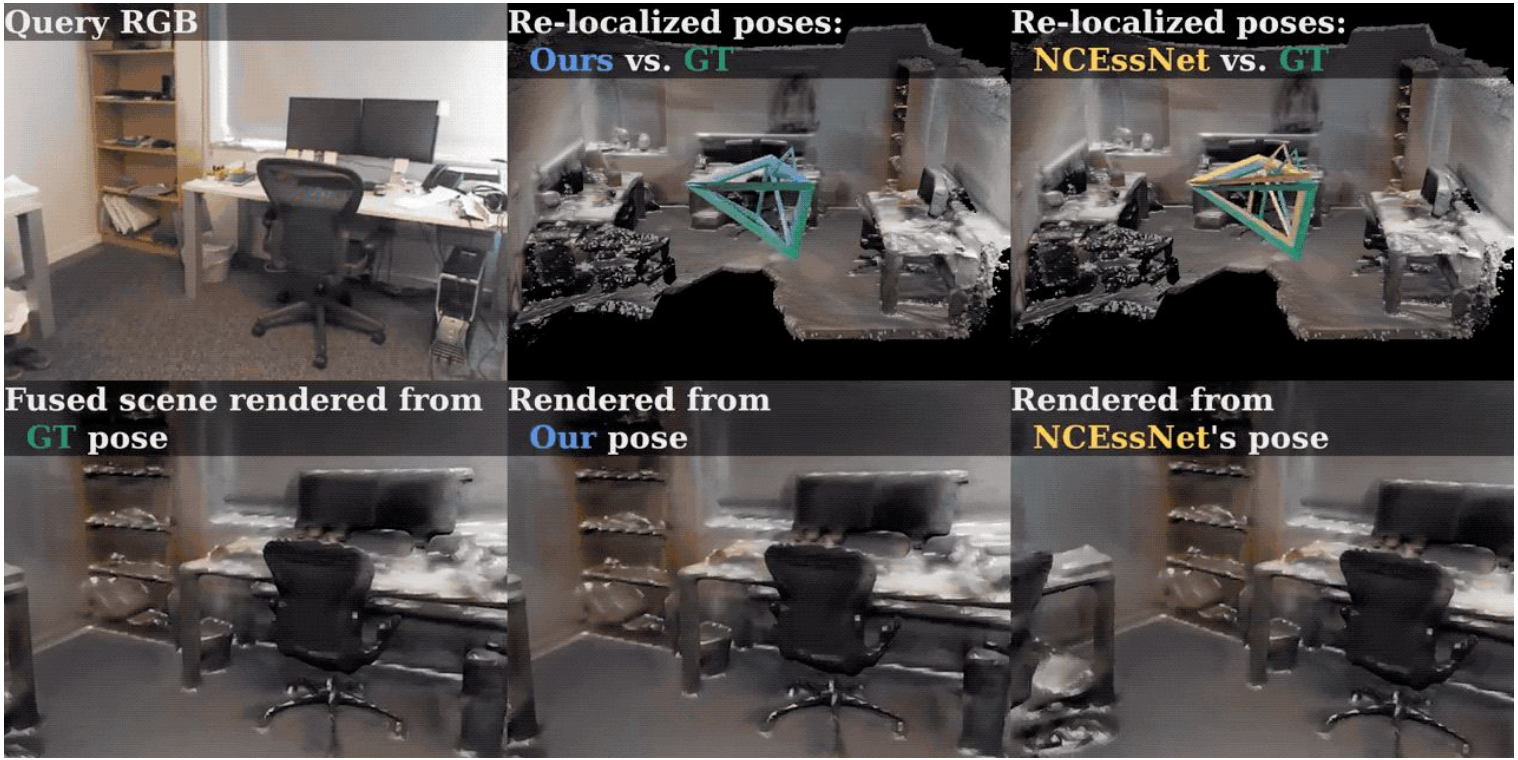
ETH zürich NIANTIC UCL

Results: Qualitative - 7-Scenes (RedKitchen)



ETH zürich NIANTIC UCL

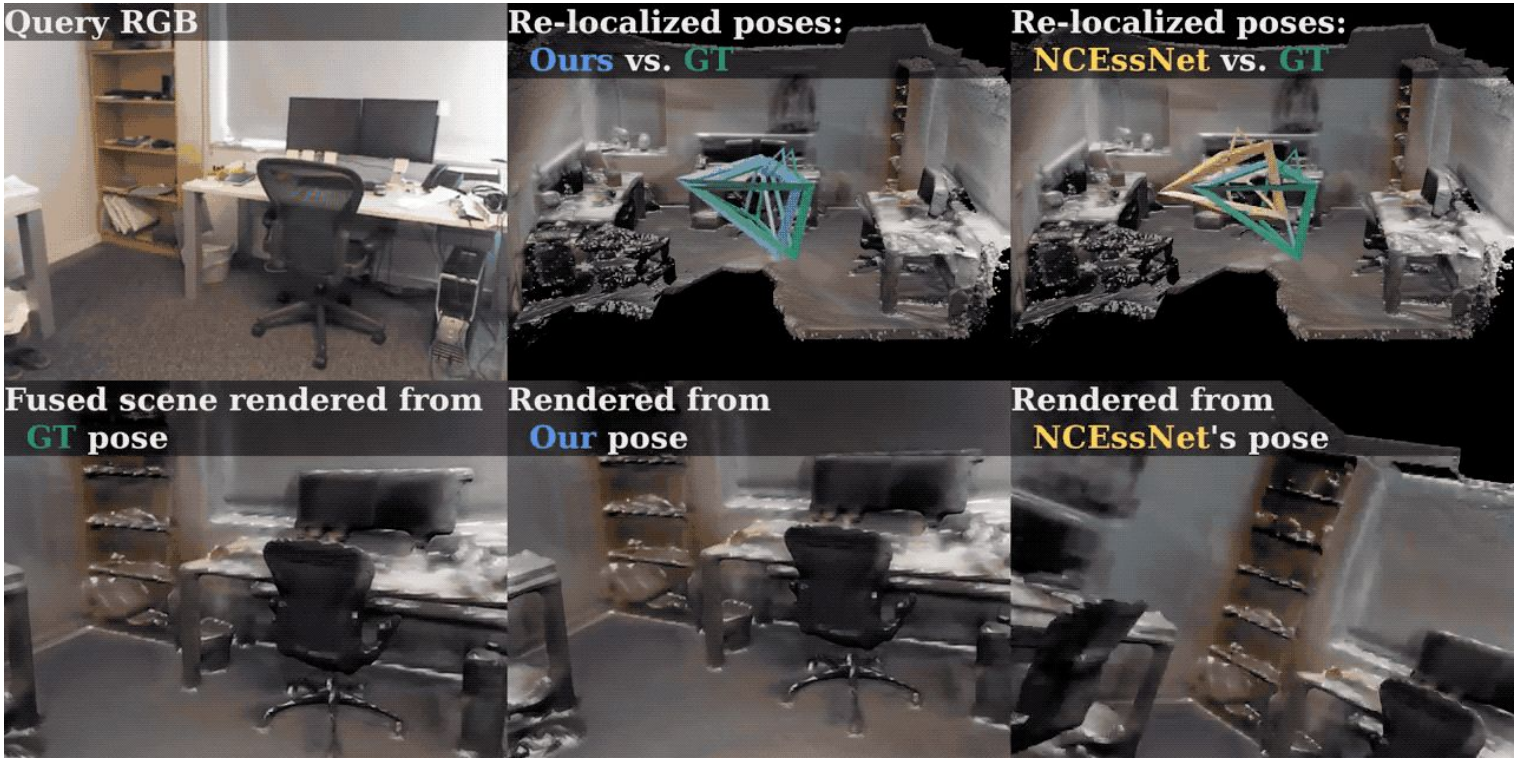
Results: Comparison to baseline (trained on 7 training sets)



[NC-EssNet] Zhou, Qunjie, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe.
"To learn or not to learn: Visual localization from essential matrices." In ICRA, 2020.

ETH zürich NIANTIC UCL

Results: Generalization to **unseen** environment (trained on **6** training sets)



[NC-EssNet] Zhou, Qunjie, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe.
"To learn or not to learn: Visual localization from essential matrices." In ICRA, 2020.

Results: Cambridge Landmarks

		# test frames	College	Hospital	Shop	Church	Court	Avg. (4)	Avg. (5)
	DSAC* [11]*	1	0.18, 0.3°	0.21, 0.4°	0.05, 0.3°	0.15, 0.5°	0.34, 0.2°	0.15, 0.4°	0.19, 0.3°
Seq.	MapNet [12]*	3	1.08, 1.9°	1.94, 3.9°	1.49, 4.2°	2.00, 4.5°	7.85, 3.8°	1.63, 3.6°	2.87, 3.7°
Seq.	GL-Net [70]*?	8	0.59, 0.7°	1.88, 2.8°	0.50, 2.9°	1.90, 3.3°	6.67, 2.8°	1.22, 2.4°	2.31, 2.5°
Image based APR	PoseNet [36]*	1	1.92, 5.4°	2.31, 5.4°	1.46, 8.1°	2.66, 8.5°	-	2.09, 6.8°	-
Image based APR	Bayesian PoseNet [34]*	1	1.74, 4.1°	2.57, 5.1°	1.25, 7.5°	2.11, 8.4°	-	1.96, 6.0°	-
Image based APR	PN learned weights [35]*	1	0.99, 1.1°	2.17, 2.9°	1.05, 4.0°	1.49, 3.4°	7.00, 3.7°	1.43, 2.9°	2.54, 3.0°
Image based APR	Geometric PoseNet [35]*	1	0.88, 1.0°	3.20, 3.3°	0.88, 3.8°	1.57, 3.3°	6.83, 3.5°	1.63, 2.7°	2.67, 3.0°
Image based APR	SVS-Pose [46]*	1	1.06, 2.8°	1.50, 4.0°	0.63, 5.7°	2.11, 8.1°	-	1.33, 5.2°	-
Image based APR	LSTM-Pose [63]*	1	0.99, 3.7°	1.51, 4.3°	1.18, 7.4°	1.52, 6.7°	-	1.30, 5.5°	-
Image based APR	GPoseNet [14]*	1	1.61, 2.3°	2.62, 3.9°	1.14, 5.7°	2.93, 6.5°	-	2.08, 4.6°	-
Image based APR	MLFBPPose [66]*	1	0.76, 1.7°	1.99, 2.9°	0.75, 5.1°	1.29, 5.0°	-	1.20, 3.7°	-
Image based APR	ADPoseNet [33]*	1	1.30, 1.7°	-	1.22, 6.7°	2.28, 4.8°	-	1.60, 4.2°	-
Image based APR	AnchorPoint [48]*?	1	0.57, **0.9°**	1.21, 2.6°	0.52, **2.3°**	**1.04**, **2.7°**	4.64, 3.4°	**0.84**, **2.1°**	1.60, 2.4°
Image based APR	AnchorPoint [48]* publ. code [49]	1	1.02_{2D}, -	0.82_{2D}, -	0.94_{2D}, -	1.02_{2D}, -	-	0.95_{2D}, -	-
IR	DenseVLAD [58]	1	2.80, 5.7°	4.01, 7.1°	1.11, 7.6°	2.31, 8.0°	-	2.56, 7.1°	-
IR	DenseVLAD+Inter [54]	1	1.48, 4.5°	2.68, 4.6°	0.90, 4.3°	1.62, 6.1°	-	1.67, 4.9°	-
RPR	EssNet [72]	1	0.76, 1.9°	1.39, 2.8°	0.84, 4.3°	1.32, 4.7°	-	1.08, 3.4°	-
RPR	NC-EssNet [72]	1	0.61, 1.6°	**0.95**, 2.7°	0.7 , 3.4°	1.12, 3.6°	-	0.85, 2.8°	-
	Ours	1	**0.48**, 1.0°	1.14, **2.5°**	**0.48**, 2.5°	1.52, 3.2°	**3.2**, **2.2°**	0.91, 2.3°	**1.37**, **2.3°**

= main competitor/ours = unreproducible baseline w/public version of the code

ETH zürich NIANTIC UCL

Results: Ablation

	# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
Baseline-1: w/o GNN	1	0.30, 14.2°	0.43, 18.2°	0.19, 18.1°	0.41, 13.9°	0.47, 15.0°	0.41, 15.1°	0.32, 16.2°	0.36, 15.8°
Baseline-2: w/o IR, rand.	1	0.14, 4.3°	0.27, 9.7°	0.15, 9.1°	0.21, 5.7°	0.23, 5.0°	0.27, 5.2°	0.28, 7.9°	0.22, 6.7°
Baseline-3: w/o IR, seq ~[70]	8	0.20, 6.8°	0.35, 12.3°	0.18, 14.2°	0.26, 7.7°	0.31, 6.9°	0.36, 7.9°	0.43, 13.1°	0.30, 9.8°
Ours	1	**0.09**, **2.7°**	**0.24**, **7.5°**	**0.13**, **8.7°**	**0.15**, **4.1°**	**0.17**, **3.5°**	**0.20**, **3.7°**	**0.23**, **6.5°**	**0.17**, **5.2°**

- **GNN**s allow efficient and effective information exchange between multiple views.

Results: Ablation

	# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
Baseline-1: w/o GNN	1	0.30, 14.2°	0.43, 18.2°	0.19, 18.1°	0.41, 13.9°	0.47, 15.0°	0.41, 15.1°	0.32, 16.2°	0.36, 15.8°
Baseline-2: w/o IR, rand.	1	0.14, 4.3°	0.27, 9.7°	0.15, 9.1°	0.21, 5.7°	0.23, 5.0°	0.27, 5.2°	0.28, 7.9°	0.22, 6.7°
Baseline-3: w/o IR, seq ~[70]	8	0.20, 6.8°	0.35, 12.3°	0.18, 14.2°	0.26, 7.7°	0.31, 6.9°	0.36, 7.9°	0.43, 13.1°	0.30, 9.8°
Ours	1	**0.09, 2.7°**	**0.24, 7.5°**	**0.13, 8.7°**	**0.15, 4.1°**	**0.17, 3.5°**	**0.20, 3.7°**	**0.23, 6.5°**	**0.17, 5.2°**

- **GNN**s allow efficient and effective information exchange between multiple views.
- GNNs work best with enough diversity and enough overlap from what **image retrieval** provides.

[GL-Net, 70] Xue, Fei, et al. "Learning multi-view camera relocalization with graph neural networks." In CVPR, 2020.

ETH zürich NIANTIC UCL

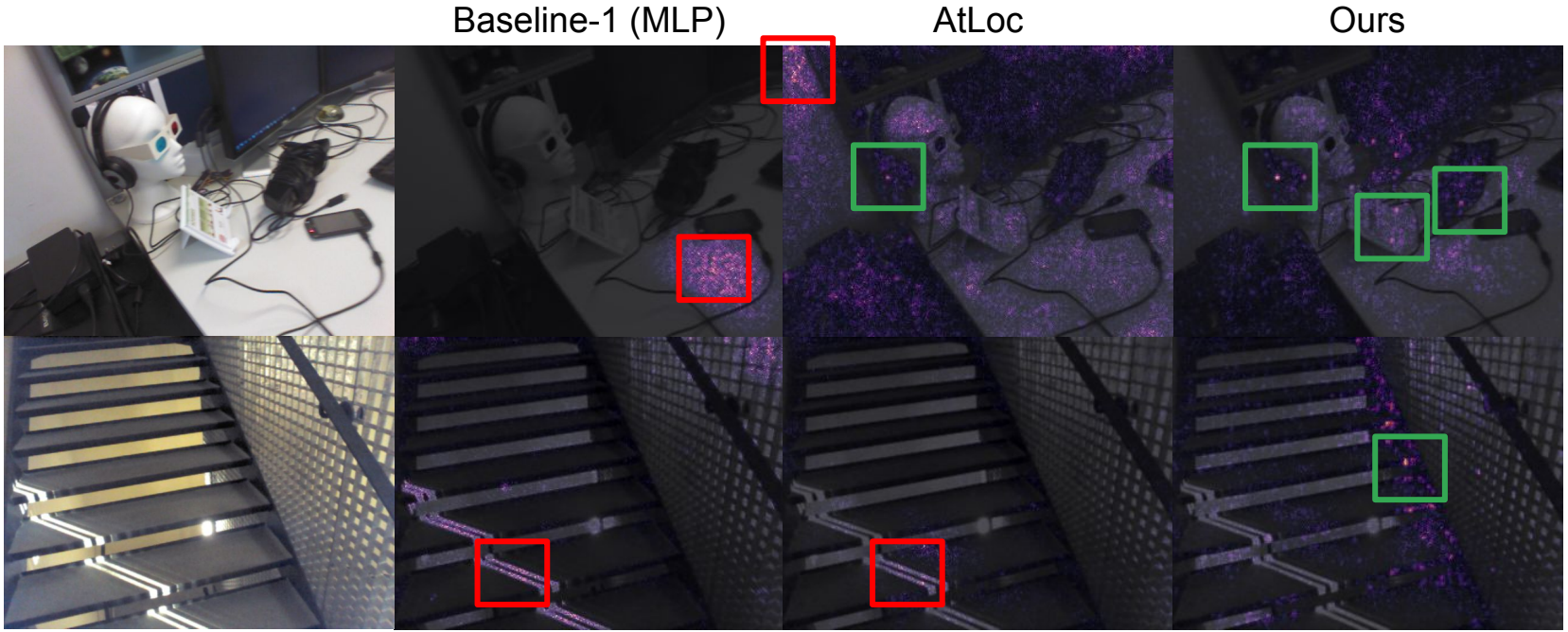
Results: Time Complexity

	Training (day)	Inference (ms)
NN-Net [4] (2x GPU)	3.1	-
CamNet [2]	15.2	157
NC-EssNet [9]	8.5	337
Ours	**0.8**	**25.7**

[NN-Net] Laskar, Zakaria, et al. "Camera relocalization by computing pairwise relative poses using convolutional neural network." In CVPR workshops, 2017.
[CamNet] Ding, Mingyu, et al. "CamNet: Coarse-to-fine retrieval for camera re-localization." In ICCV, 2019.
[NC-EssNet] Zhou, Qunjie, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. "To learn or not to learn: Visual localization from essential matrices." In ICRA, 2020.

Results: Saliency Visualization

A saliency map offers a visualization of pixels that contribute the most to predictions.



- Our model learns more robust/less scene specific features

[AtLoc] Wang, Bing, et al. "Atloc: Attention guided camera localization." In AAAI, 2020.

ETH zürich NIANTIC UCL

Contributions

1. First to apply GNNs for relative pose re-localization.

ETH zürich NIANTIC UCL

Contributions

1. First to apply GNNs for relative pose re-localization.

2. New reproducible SOTA baseline for learning-based RPR approaches.

ETH zürich NIANTIC UCL

Contributions

1. First to apply GNNs for relative pose re-localization.
2. New reproducible SOTA baseline for learning-based RPR approaches.
3. Significantly reduces RPR methods' time complexity.

ETH zürich NIANTIC UCL

Lessons Learned

- Communication of multiple views
  - **GNN**s allow efficient info exchange between multiple views
- Diversity vs. Overlap
  - GNNs work best with enough diversity and enough overlap from what **image retrieval** provides
- **Attention** in message passing
  - Important to focus on different features for different view pairs
- Training with **multiple scenes** allows
  - Learning more robust/less scene specific features

ArXiv link

Code: https://github.com/nianticlabs/relpose-gnn

Contact:

- ozgur.turkoglu@geod.baug.ethz.ch
- aron@nianticlabs.com