

# Visual Camera Re-Localization

## Using Graph Neural Networks and Relative Pose Supervision



Mehmet Özgür Türkoğlu<sup>1\*</sup>



Eric Brachmann<sup>2</sup>



Konrad Schindler<sup>1</sup>



Gabriel J. Brostow<sup>2,3</sup>



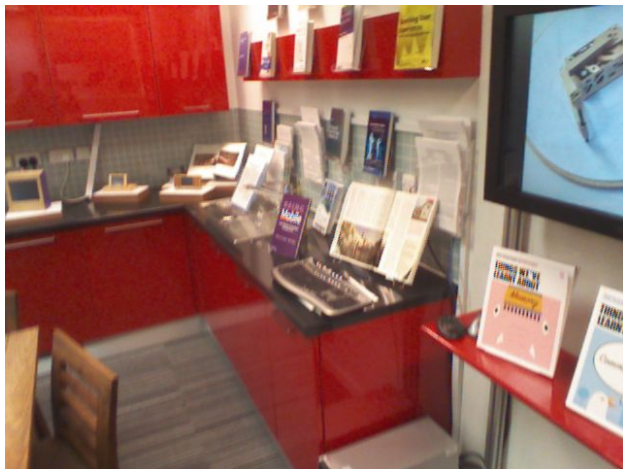
Áron Monszpart<sup>2</sup>

<sup>1</sup>ETH Zurich

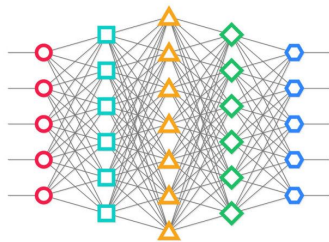
<sup>2</sup>Niantic

<sup>3</sup>University College London

# Visual Camera Re-localization



Query  
image



Algorithm e.g.  
CNN-based pose  
regressor



6-DoF camera  
pose

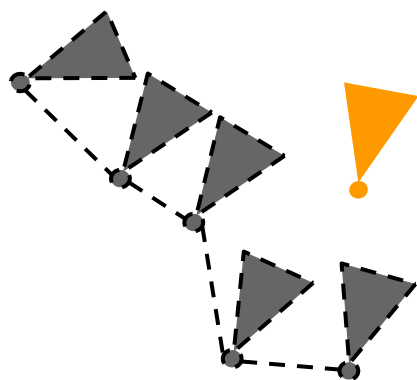
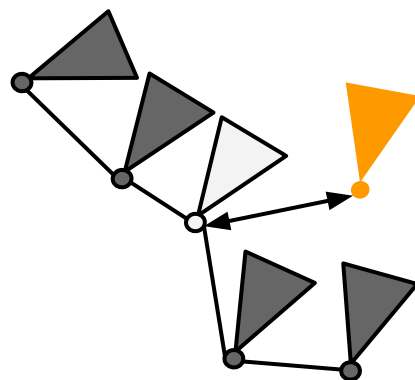
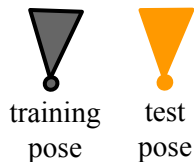
## Motivation

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

- Intrinsic 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.*, PoseNetRPR *e.g.*, RelocNet

--- = only during training

— = during both training/test

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

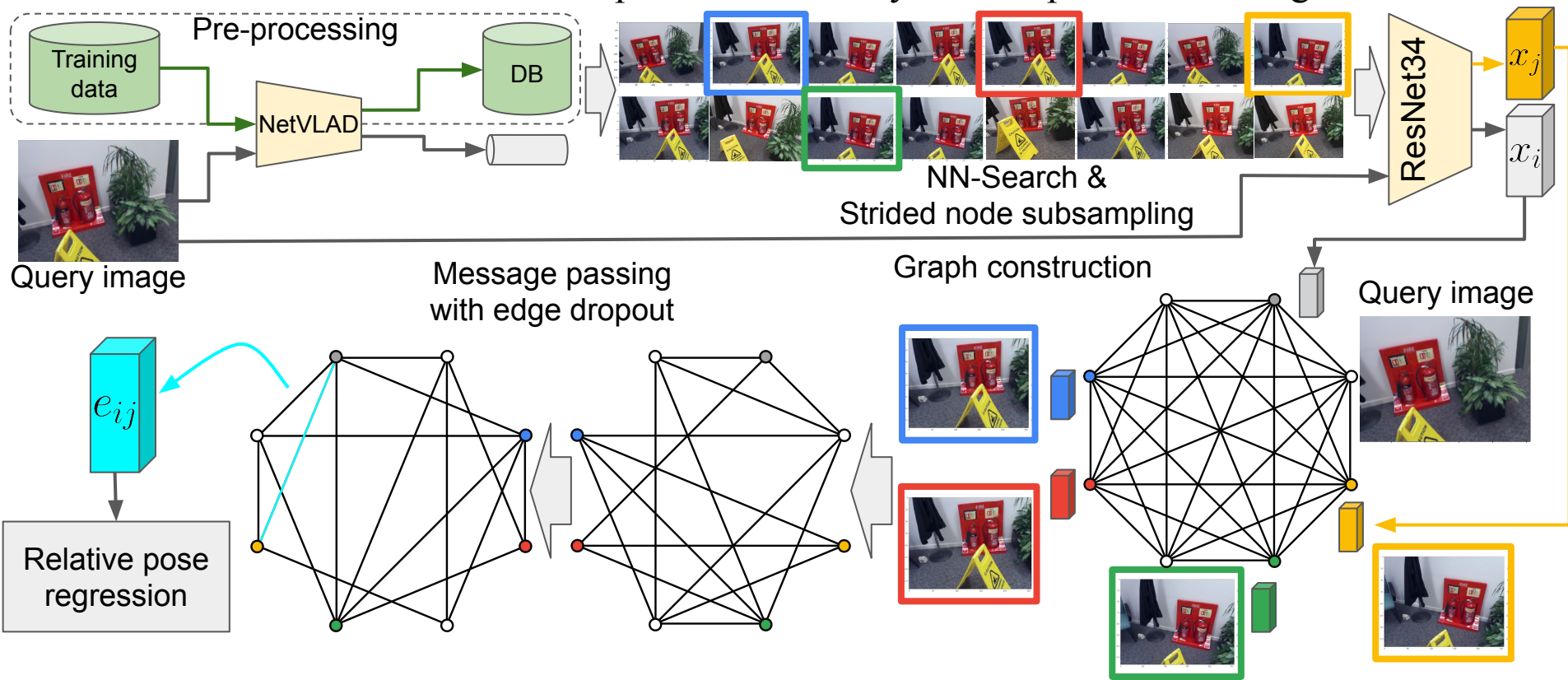
😞=weak    😐=okay    😎=promising    💪= strong

**[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.

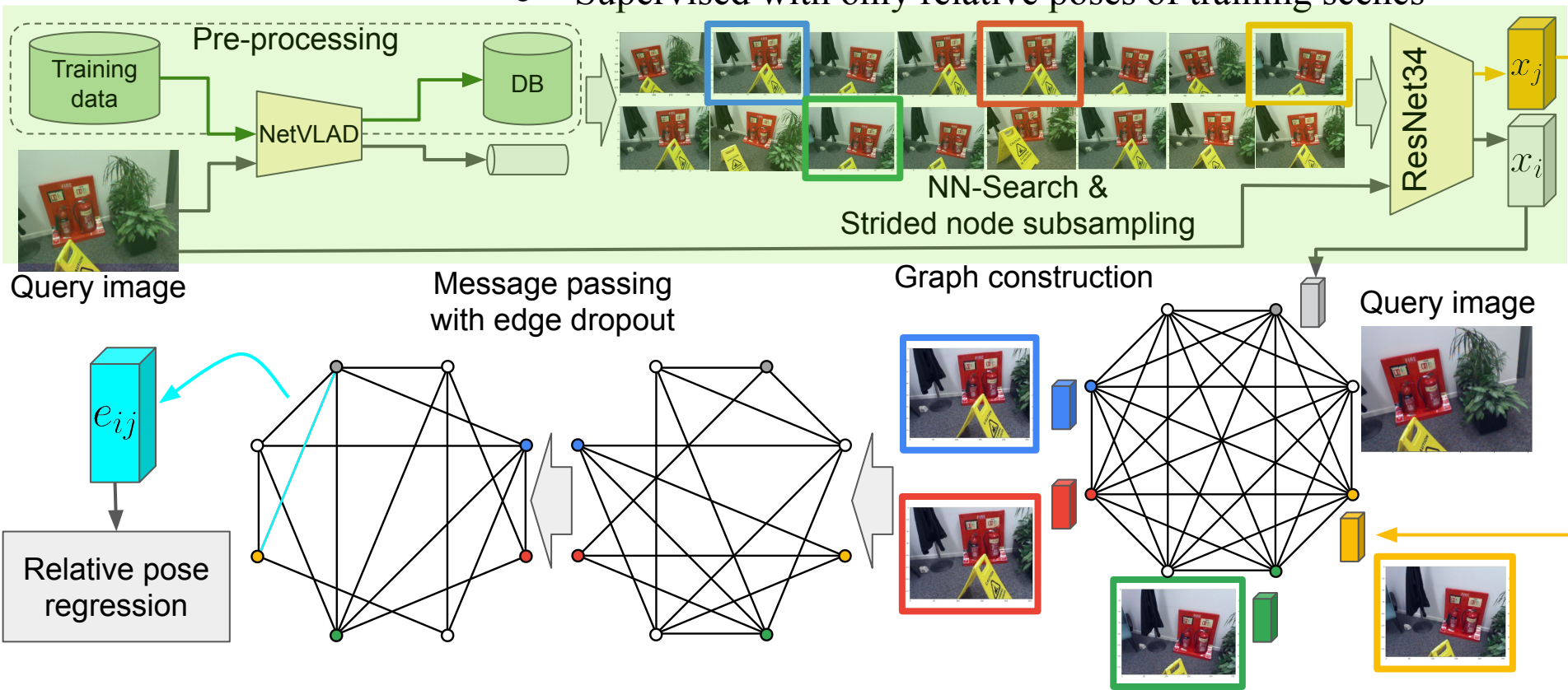
## Method Overview

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



# Method Overview

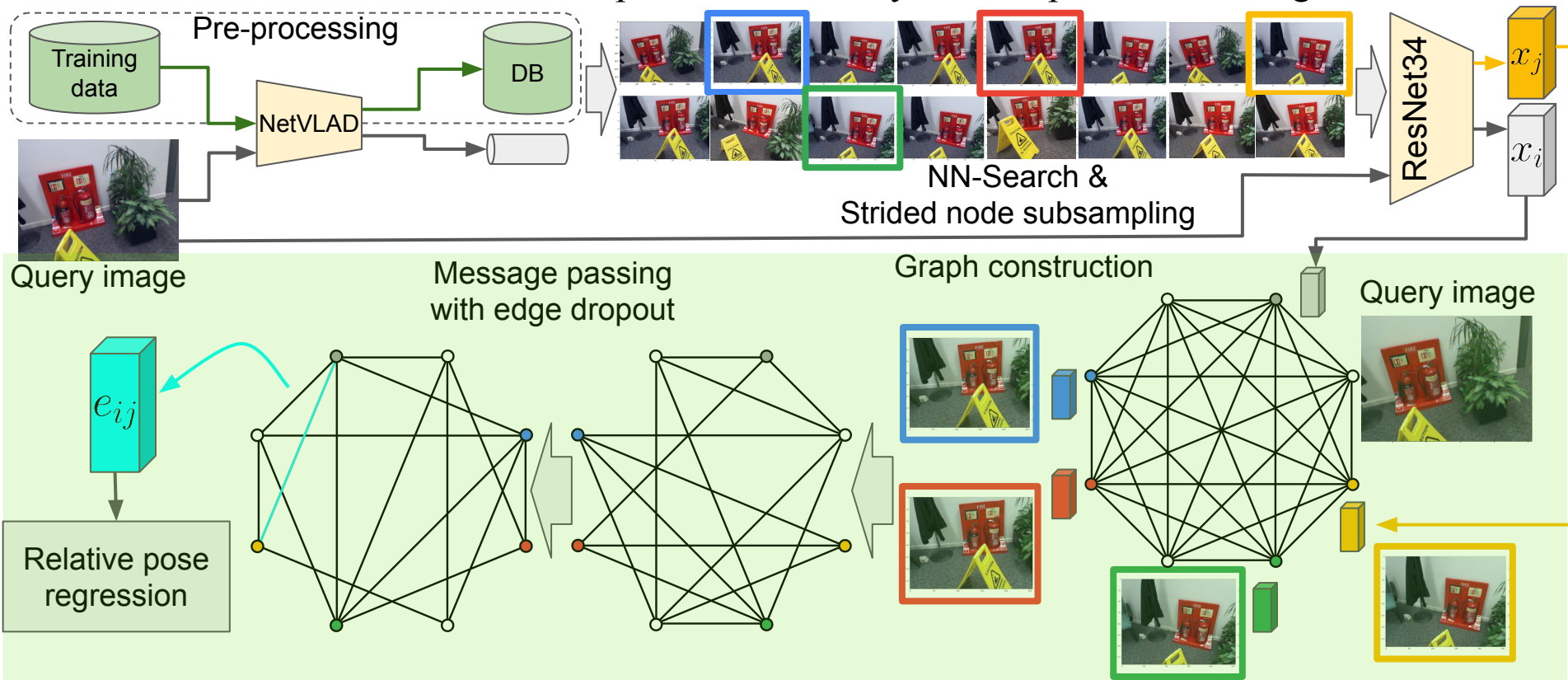
- Image retrieval + GNN
- Supervised with only relative poses of training scenes





## Method Overview

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



## Results: 7-Scenes

	# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.	
Seq. based	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°
	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]* <sub>?</sub>	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]* <sub>?</sub>	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] <sub>?</sub>	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.	
Self. based	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°
	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]* <sub>7</sub>	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]* <sub>7</sub>	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] <sub>7</sub>	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

## Results: Time Complexity

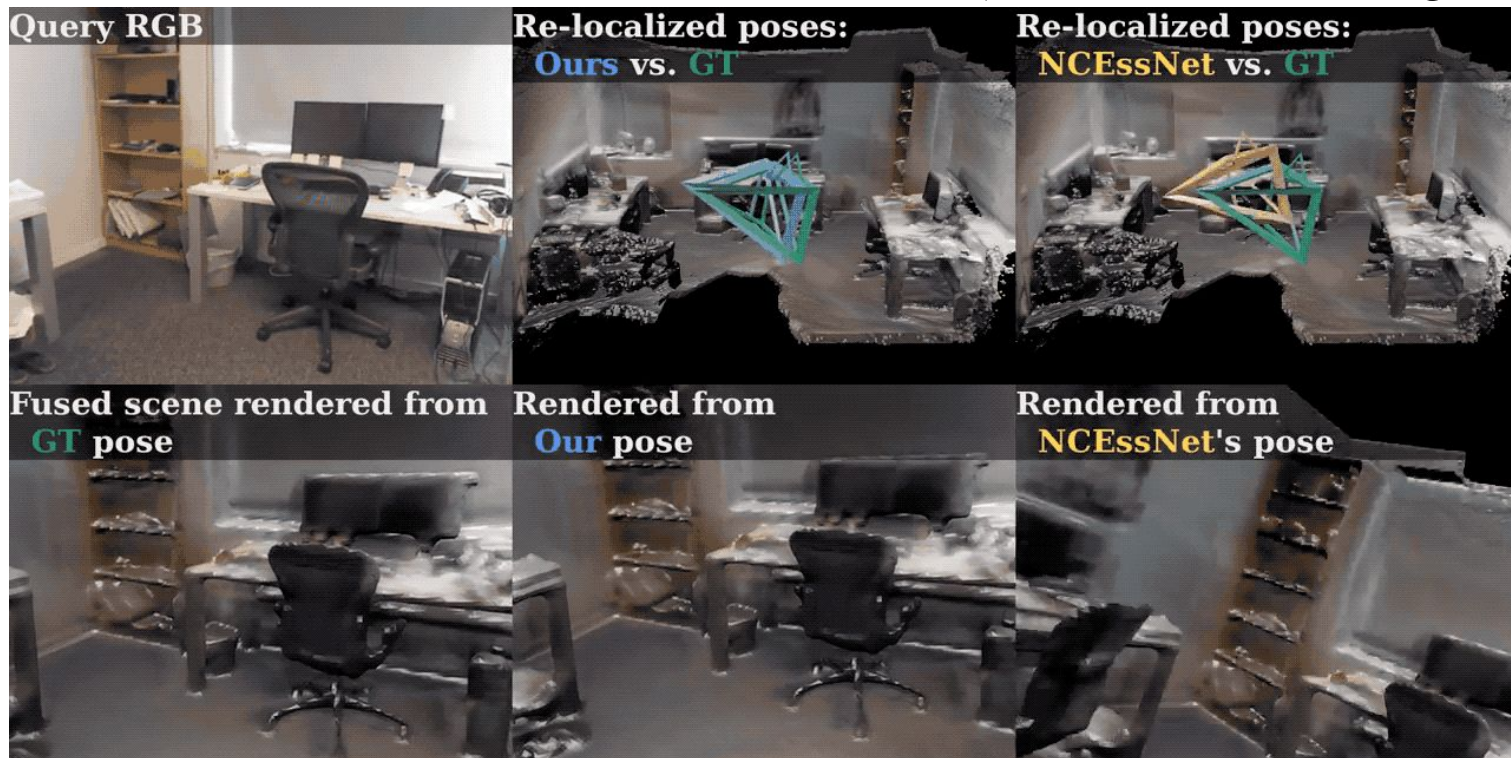
	Training (day)	Inference (ms)
NN-Net [4] (2x GPU)	3.1	-
CamNet [2]	15.2	157
NC-EssNet [9]	8.5	337
Ours	<b>0.8</b>	<b>25.7</b>

[**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: 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.