ETH zürich · NIANTIC · UCL

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

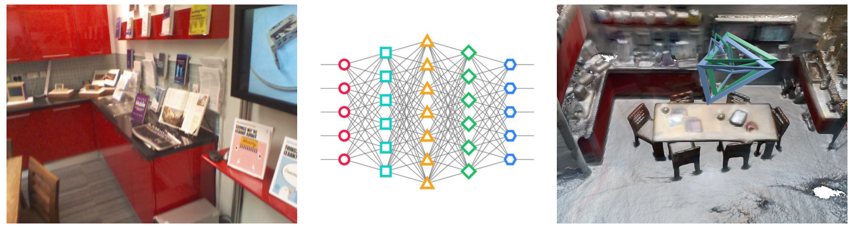
3DV 2021

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

1 Visual Re-localization

means using a single image as input to estimate the camera's location and orientation relative to a pre-recorded environment.



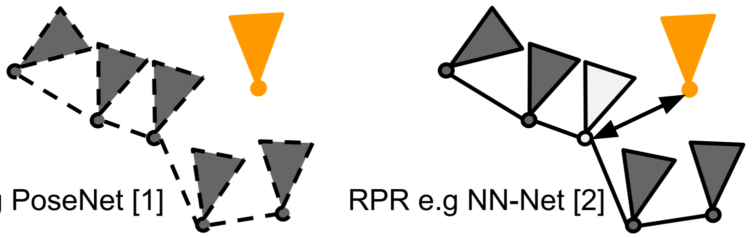
Query image · Algorithm · 6-DoF camera pose

2 Motivation for Learning-based Approaches

Structure-based methods [3] achieve SOTA.
So why look beyond structure-based methods?

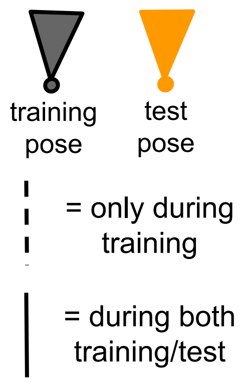
- Intrinsics are often not available or reliable
- Geometric optimization is costly
- Work best for scenes with easy-to track feature points

3 Deep **A**bsolute vs **R**elative **P**ose **R**egression



APR e.g PoseNet [1] · RPR e.g NN-Net [2]

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



training pose · test pose

= only during training

= during both training/test

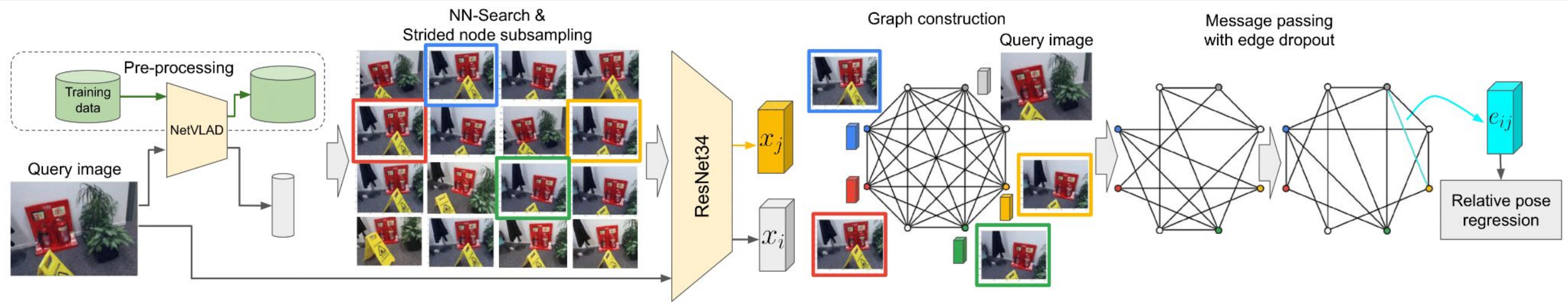
😩=weak 😐=okay 😎=promising 💪= strong

4 Method

Our method consists:

1. Visual Encoding
2. Image retrieval
3. Graph construction
4. Message passing

🤘 Trained with only relative pose supervision.



5 Experimental Results

	Method	# 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, -
Seq. based	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°
Seq. based	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°
Seq. based	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
Image based APR	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°
IR	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°
RPR	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°
RPR	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°
RPR	EssNet [72] reprod.	1	-	-	-	-	-	-	0.32, 9.8°	-
RPR	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°
RPR	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 (DenseVLAD, DenseVLAD+Inter, NN-Net, RelocNet, EssNet, EssNet reprod., NC-EssNet, NC-EssNet reprod., Ours) · = baseline unreproducible w/public version of the code (GL-Net, AnchorPoint, CamNet)

Ablation

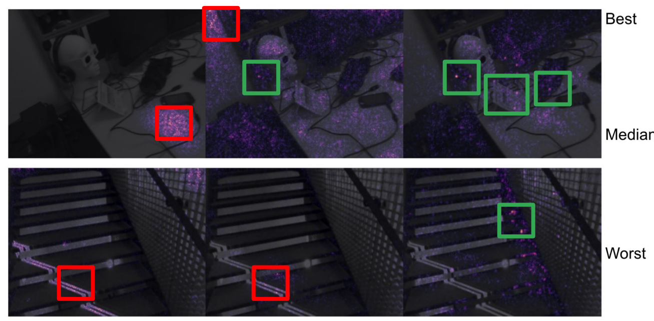
	# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
Baseline-1: w/o GNN	1	0.30, 14.2°	0.43, 18.2°	0.19, 18.1°	0.41, 13.9°	0.47, 15.0°	0.41, 15.1°	0.32, 16.2°	0.36, 15.8°
Baseline-2: w/o IR, rand.	1	0.14, 4.3°	0.27, 9.7°	0.15, 9.1°	0.21, 5.7°	0.23, 5.0°	0.27, 5.2°	0.28, 7.9°	0.22, 6.7°
Baseline-3: 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°**

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**

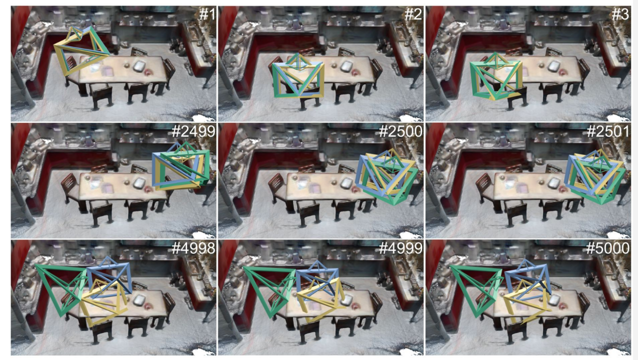
Saliency Visualization

Baseline-1 (MLP) · AtLoc · Ours



Qualitative

Ours · NC-EssNet · GT



6 Contributions

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

7 Lessons Learned

- Communication of multiple views
  - **GNNs** 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
  - Learn more robust/less scene specific features

References

[1] Kendall et al. "Posenet: A convolutional network for real-time 6-dof camera relocalization." In ICCV, 2015.
[2] Laskar et al. "Camera relocalization by computing pairwise relative poses using convolutional neural network." In ICCV Workshops, 2017.
[3] Sattler et al. "Understanding the limitations of CNN-based absolute camera pose regression." In CVPR, 2019