The NeRFect Match: Exploring NeRF Features for Visual Localization

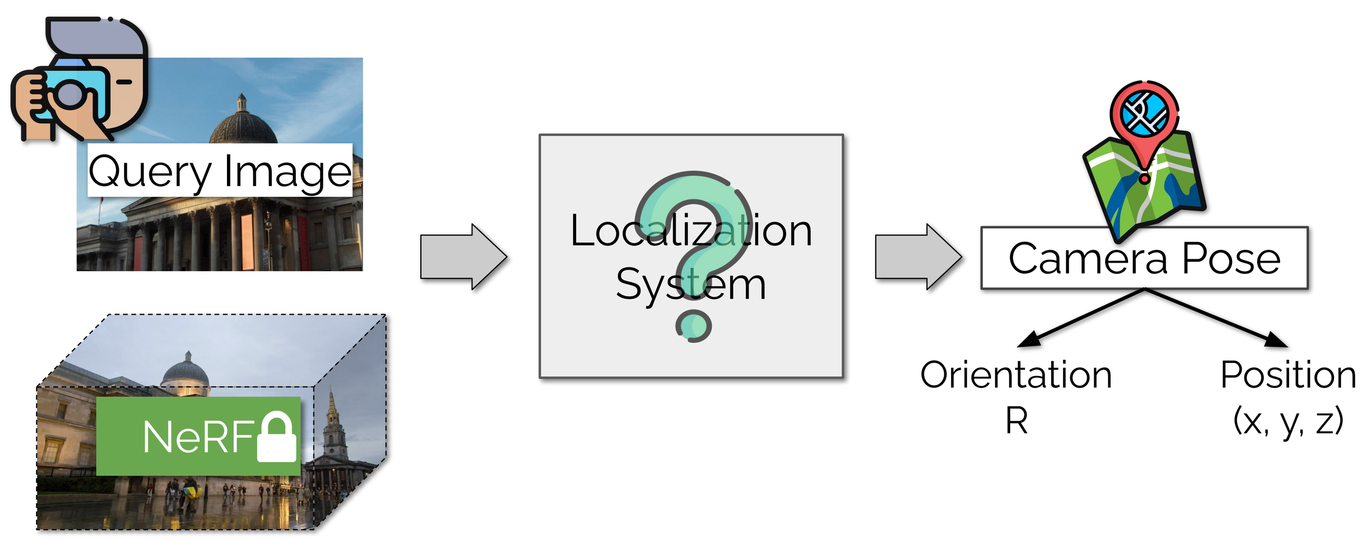
Qunjie Zhou [1] Maxim Maximov [1,2] Or Litany [1,3] Laura Leal-Taixé [1]

[1] NVIDIA [2] TU Munich [3] Technion

Introduction

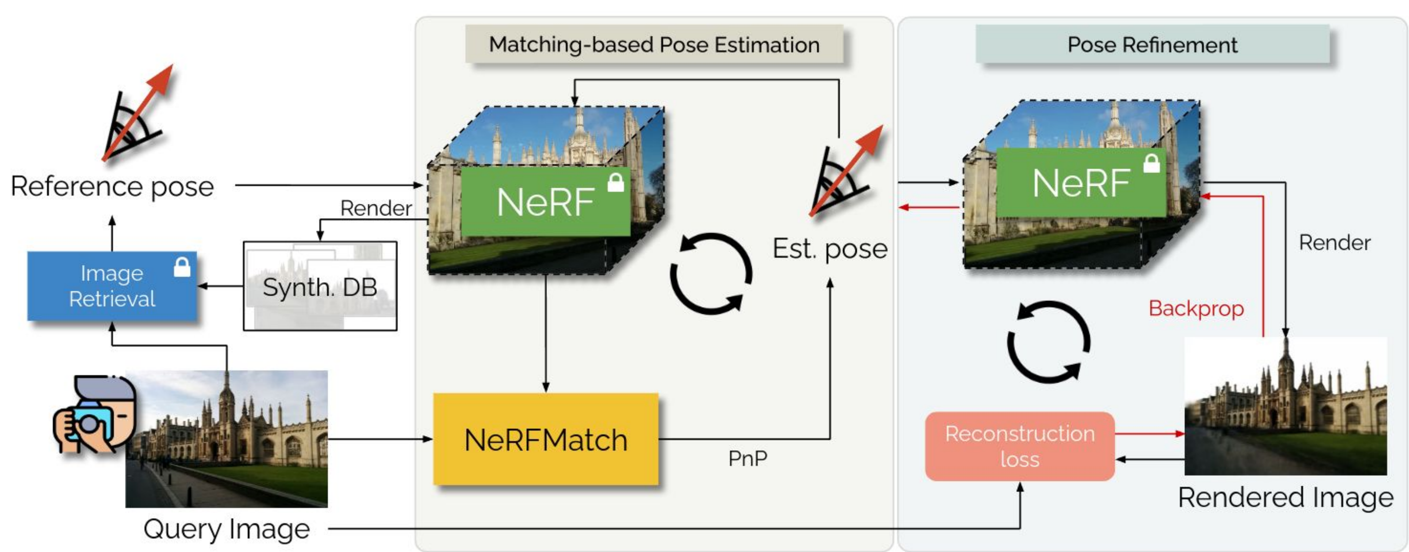
Motivation

Given a RGB query image, our goal is to localize its camera pose w.r.t a 3D scene. We propose to use NeRF as a ***compact*** and ***interpretable*** dense scene representation for visual localization.

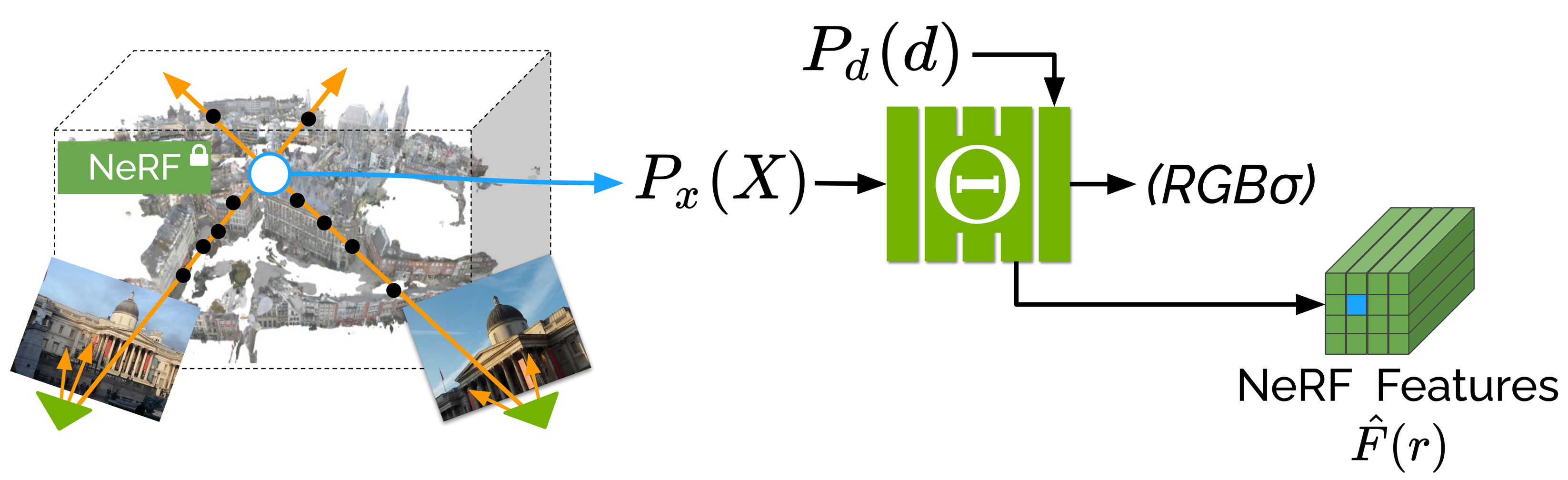


NeRF-based Localization

- Our hierarchical NeRF-based localization pipeline directly estimates 2d-3d correspondences between a query image and the scene representation ***without*** **keeping an expensive 3D point cloud** of the scene.
- Compared to other NeRF-based localization, we use NeRF as the **primary** scene representation ***without*** **re-training or modifications**.

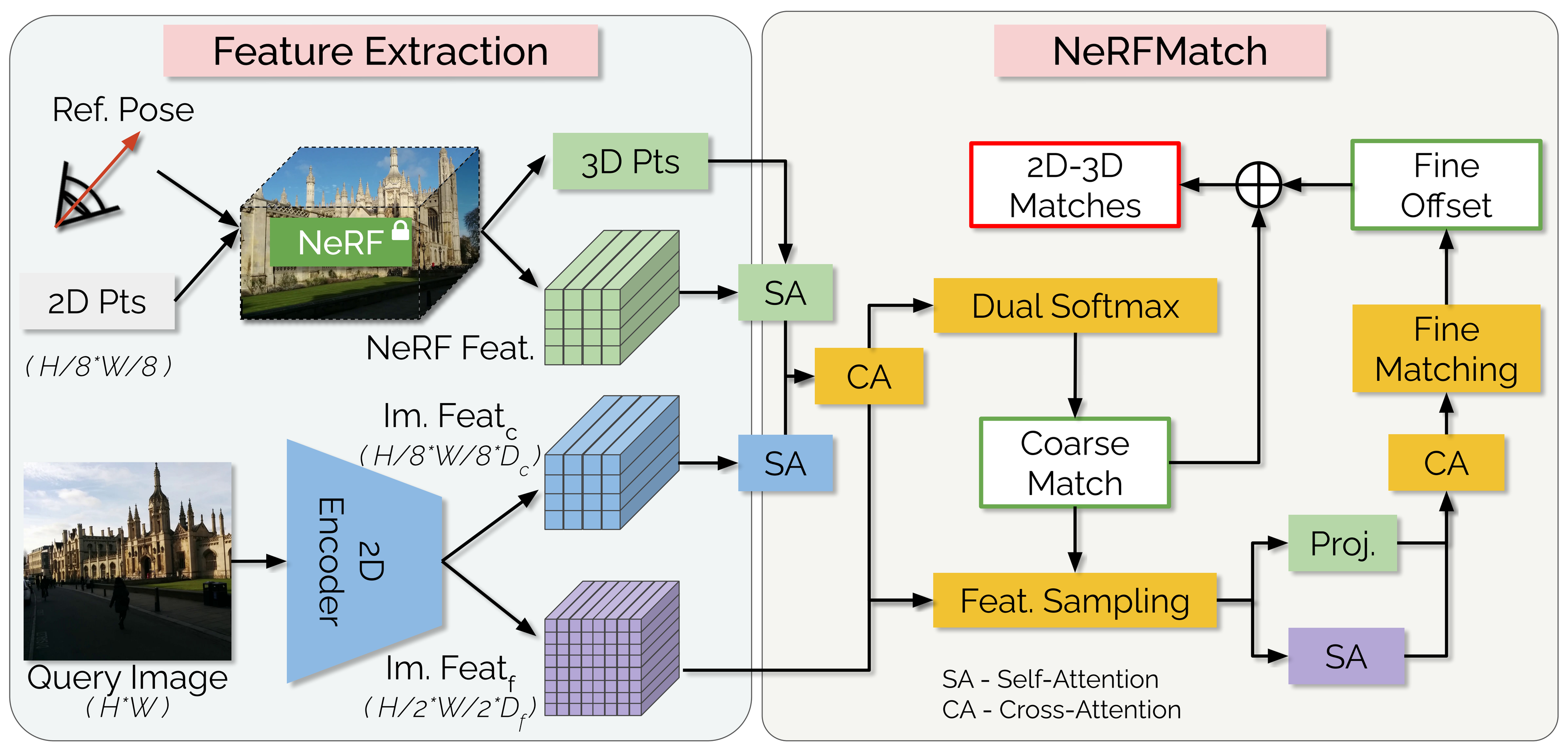


NeRFMatch Model



NeRF Feature Rendering

A sampled 3D point X is passed into a **pre-trained NeRF** model to extract 3D features from an internal MLP layer. Volumetric rendering aggregates features along a ray to obtain the NeRF descriptor for a 3D **surface** point.



NeRFMatch Architecture

- **NeRF features** rendered from a given reference pose
- **Image features** extracted using pre-trained ConvFormer encoder
- **Coarse-level matching** for 3D point-to-image patch (8x8) correspondences
- **Fine-level matching** for 3D point-to-image patch (2x2) correspondences

Comparison to SOTA

	Method	Scene Repres.	Cambridge Landmarks - Outdoor: Kings	Hospital	Shop	StMary	Court	Avg.Med ↓
End-to-End	MS-Trans. [56]	APR Net.	83/1.5	181/2.4	86/3.1	162/4	-	-
	DFNet [17]	APR Net.	73/2.4	200 /3	67/2.2	137/4	-	-
	LENS [44]	APR Net.	33/0.5	44/0.9	27/1.6	53/1.6	-	-
	NeFeS [16]	APR+NeRF	37/0.6	55/0.9	14/0.5	32/1	-	-
	DSAC* [10]	SCR Net.	15/0.3	21/0.4	5/0.3	13/0.4	49/0.3	20.6/0.3
	HACNet [36]	SCR Net.	18/0.3	19/**0.3**	6/0.3	9/0.3	28/0.2	16/0.3
	ACE [6]	SCR Net.	28/0.4	31/0.6	5/0.3	18/0.6	43/0.2	25/0.4
Hierachical	SANet [72]	3D+RGB	32/0.5	32/0.5	10/0.5	16/0.6	328/2.0	83.6/0.8
	DSM [62]	SCR Net.	19/0.4	24/0.4	7/0.4	12/0.4	44/0.2	21.2/0.4
	NeuMap [63]	SCode+RGB	14/**0.2**	19/0.4	6/0.3	17/0.5	**6/0.1**	12.4/0.3
	InLoc [60]	3D+RGB	46/0.8	48/1.0	11/0.5	18/0.6	120/0.6	48.6/0.7
	HLoc [51]	3D+RGB	12/**0.2**	**15/0.3**	**4/0.2**	**7/0.2**	16/**0.1**	10.8/**0.2**
	PixLoc [53]	3D+RGB	14/**0.2**	16/**0.3**	5/**0.2**	10/0.3	30/**0.1**	15/**0.2**
	CrossFire [43]	NeRF+RGB	47/0.7	43/0.7	20/1.2	39/1.4	-	-
	NeRFLoc [38]	NeRF+RGBD	**11/0.2**	18/0.4	**4/0.2**	**7/0.2**	25/**0.1**	13/**0.2**
	NeRFMatch-Mini	NeRF+RGB	19.0/0.3	30.2/0.6	10.3/0.5	11.3/0.4	29.1/0.2	20.0/0.4
	NeRFMatch	NeRF+RGB	13.0/**0.2**	19.4/0.4	8.5/0.4	7.9/0.3	17.5/**0.1**	13.3/0.3
	NeRFMatch	NeRF	12.7/**0.2**	20.7/0.4	8.7/0.4	11.3/0.4	19.5/**0.1**	14.6/0.3

Method	Scene Repres.	7-Scenes - SfM Poses - Indoor: Chess	Fire	Heads	Office	Pump.	Kitchen	Stairs	Avg.Med↓	Avg.Recall↑
MS-Trans. [56]	APR Net.	11/6.4	23/11.5	13/13	18/8.1	17/8.4	16/8.9	29/10.3	18.1/9.5	-
DFNet [17]	APR Net.	3/1.1	6/2.3	4/2.3	6/1.5	7/1.9	7/1.7	12/2.6	6.4/1.9	-
NeFeS [16]	APR+NeRF	2/0.8	2/0.8	2/1.4	2/0.6	2/0.6	2/0.6	5/1.3	2.4/0.9	-
DSAC* [10]	SCR Net.	0.5/0.2	0.8/0.3	0.5/0.3	1.2/0.3	1.2/0.3	0.7/0.2	2.7/0.8	1.1/0.3	**97.8**
ACE [6]	SCR Net.	0.7/0.5	0.6/0.9	0.5/ 0.5	1.2/0.5	1.1/0.2	0.9/0.5	2.8/1.0	1.1/0.6	97.1
DVLAD+R2D2 [48,64]	3D+RGB	**0.4/0.1**	**0.5/0.2**	**0.4/0.2**	**0.7/0.2**	**0.6/0.1**	**0.4/0.1**	**2.4/0.7**	**0.8/0.2**	95.7
HLoc [51]	3D+RGB	0.8/**0.1**	0.9/**0.2**	0.6/0.3	1.2/**0.2**	1.4/0.2	1.1/**0.1**	2.9/0.8	1.3/0.3	95.7
NeRFMatch-Mini	NeRF+RGB	1.6/0.5	1.5/0.6	1.4/0.9	3.6/1.0	3.5/0.9	1.7/0.5	8.5/2.1	3.1/0.9	74.4
NeRFMatch	NeRF+RGB	0.9/0.3	1.1/0.4	1.4/1.0	3.0/0.8	2.2/0.6	1.0/0.3	9.0/1.5	2.7/0.7	78.2
NeRFMatch	NeRF	0.9/0.3	1.1/0.4	1.5/1.0	3.0/0.8	2.2/0.6	1.0/0.3	10.1/1.7	2.8/0.7	78.4

Insights

- Competitive outdoor localization on Cambridge Landmarks where we scale better than SCR / APR methods for larger scenes.
- Noticeable indoor performance gap on 7-Scenes due to the lack of accurate depth prediction needed for precise *centimeter-level* supervision.
- Slight performance decrease when switching to synthesized images due to the domain gap between rendered and real images.

Ablation Study

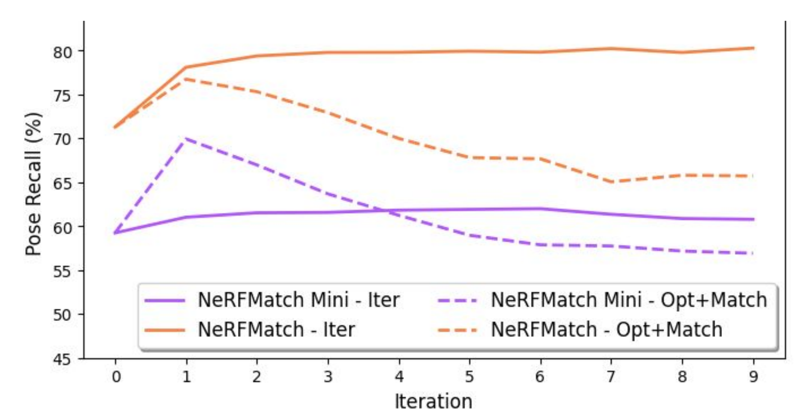
NeRF Features

Metrics	Pt3D	Pe3D	f^1	f^2	f^3	f^4	f^5	f^6	f^7
Med. Translation (cm, ↓)	458.0	34.3	28.7	28.4	**27.9**	28.3	28.3	30.2	61.3
Med. Rotation (°, ↓)	6.5	0.6	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**	1.3
Localize Recall. (%, ↑)	0.7	51.4	58.6	**59.4**	**59.2**	56.9	57.7	53.0	38.8

- Raw 3D coordinate features do not yield accurate results, yet performance improves significantly by encoding it with a positional encoding layer.
- NeRF-encoded features are **generally more effective** for matching with 2D image features, with the middle *3rd* layer showing the best results.

Pose Refinement

Model / Metrics	Best Refinement	No Refinement (top−1)	Refined (top−1)	Refined (top−10)
		Avg.Med (cm/°) ↓ / Avg.Recall (%) ↑		
NeRFMatch-Mini	Opt+Match	27.9/0.5/59.2	20.5/0.4/70.9	20.5/0.4/70.9
NeRFMatch	Iter.	16.5/0.3/71.3	14.2/0.3/78.2	13.3/0.3/80.8



- Different matching model both benefit from refinement, yet favour different strategies based on initial accuracy.
- There is a clear limit on improvement from refinement, suggesting that an **accurate initial estimation** is still the key to the accuracy.

Conclusions

- **Initial steps** towards leveraging **NeRF as the primary representation** for the task of visual localization.
- Thorough studies conducted on architectural design, 3D feature extraction and training strategies, we demonstrate **inherent capability of NeRF features to effectively support 2D-3D matching**, resulting in competitive outdoor visual localization.
- Our model directly **benefits from more accurate and efficient NeRF models** for improved localization performance..

Refer to our paper for more details !

Website