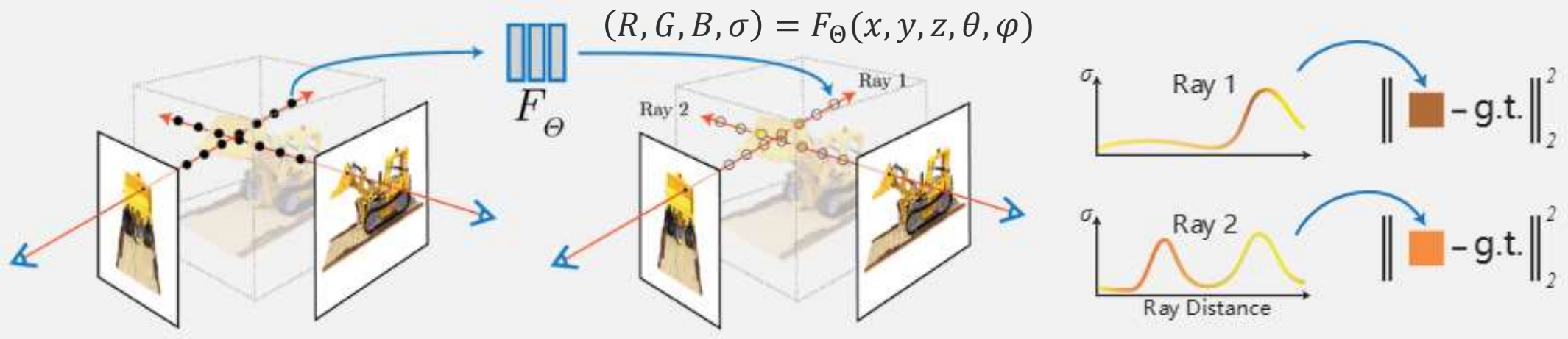
清华大学车辆与运载学院
SCHOOL OF VEHICLE AND MOBILITY TSINGHUA UNIVERSITY

NeRF-based Localization

苗津毓

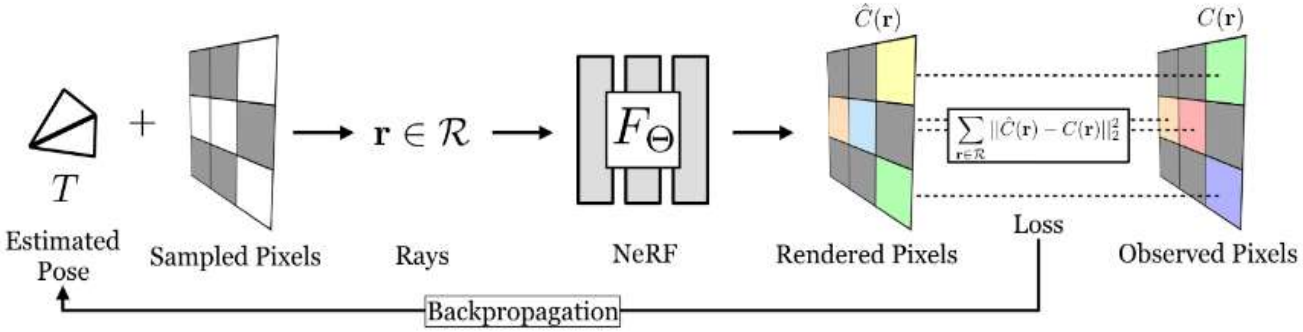
2023.11.10

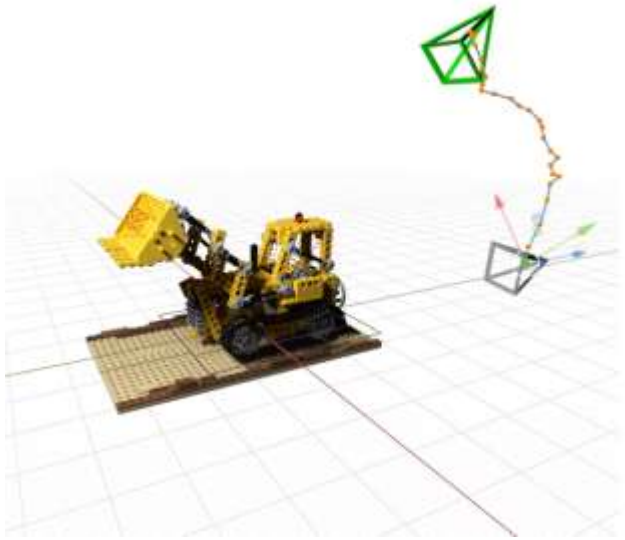
NeRF



1. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R. (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: Proceedings of European Conference on Computer Vision

iNeRF https://github.com/yenchenlin/iNeRF-public





NeRF训练的loss function

$$\mathcal{L} = \sum_{\mathbf{r}\in\mathcal{R}} ||\hat{C}(\mathbf{r}) - C(\mathbf{r})||_2^2$$

iNeRF优化位姿的loss function

$$\hat{T} = \underset{T\in SE(3)}{\text{argmin}}\ \mathcal{L}(T \mid I, \Theta)$$

iNeRF定位表现

Methods	Rotation (°) Mean	Rotation (°) Median	Translation (°) Mean	Translation (°) Median	Outlier %
SuperGlue [30]	9.27	6.22	18.2	5.4	33.3
Ours	4.39	2.01	4.81	1.77	8.7

1. L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola and T. -Y. Lin, "iNeRF: Inverting Neural Radiance Fields for Pose Estimation," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp. 1323-1330, doi: 10.1109/IROS51168.2021.9636708.

iNeRF

https://github.com/yenchenlin/iNeRF-public

A. Gradient-based SE(3) Optimization

iNeRF优化位姿的loss function:

$$\hat{T} = \underset{T \in \mathrm{SE}(3)}{\operatorname{argmin}} \mathcal{L}(T \mid I, \Theta)$$

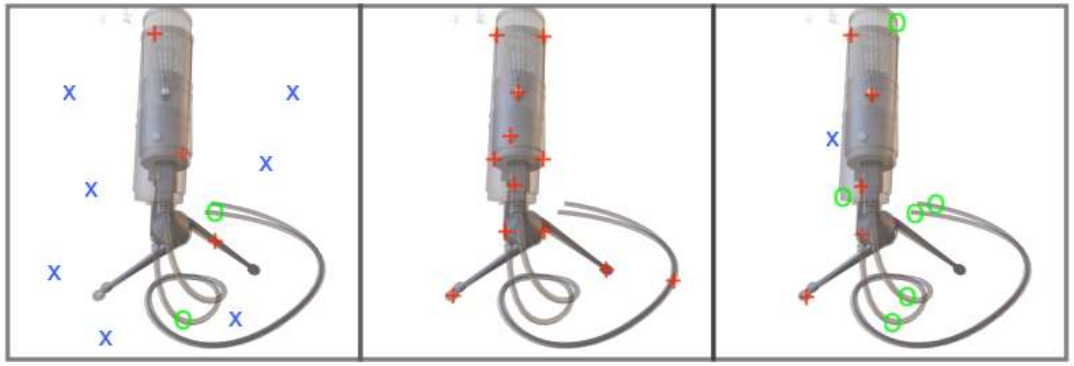
- 损失函数在SE(3)上是非凸的
- 难以保证预测的位姿T还属于SE(3)

$$\hat{T}_i = e^{[\mathcal{S}_i]\theta_i} \hat{T}_0 \,,$$

$$\text{where} \quad e^{[\mathcal{S}]\theta} = \begin{bmatrix} e^{[\omega]\theta} & K(\mathcal{S}, \theta) \\ 0 & 1 \end{bmatrix}$$

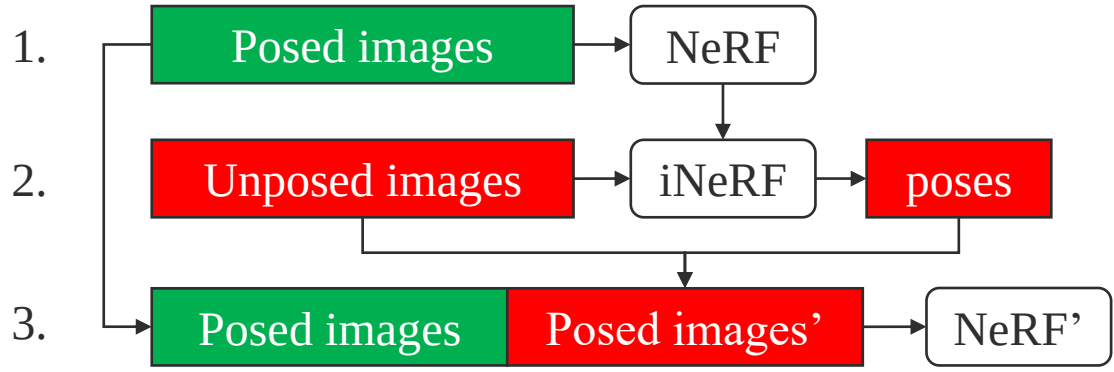
$$\widehat{\mathcal{S}\theta} = \underset{\mathcal{S}\theta \in \mathbb{R}^6}{\operatorname{argmin}} \mathcal{L}(e^{[\mathcal{S}]\theta} T_0 \mid I, \Theta).$$

B. Sampling Rays



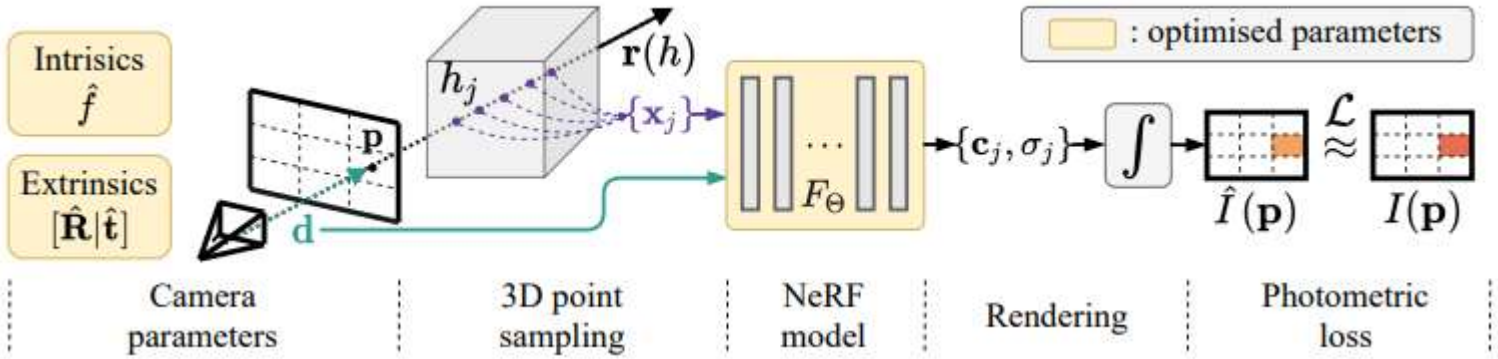
Random　　Interest Point　　Interest Region

C. Self-supervised NeRF with iNeRF



1. L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola and T. -Y. Lin, "iNeRF: Inverting Neural Radiance Fields for Pose Estimation," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp. 1323-1330, doi: 10.1109/IROS51168.2021.9636708.

NeRFmm https://github.com/ActiveVisionLab/nerfmm



已知相机内外参的情况下，NeRF训练的loss function：

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(\hat{\mathcal{I}}|\mathcal{I}, \Pi).$$

未知相机内外参的情况下，NeRF训练的loss function：

$$\Theta^*, \Pi^* = \arg\min_{\Theta,\Pi} \mathcal{L}(\hat{\mathcal{I}}, \hat{\Pi}|\mathcal{I}),$$

实际上，仅需估计焦距f和外参R, t，R用角轴表示

Sampling along the ray

$$\hat{\mathbf{r}}_{i,p}(h) = \hat{\mathbf{o}}_i + h\hat{\mathbf{d}}_{i,p}$$

$$\hat{\mathbf{o}}_i = \hat{\mathbf{t}}_i$$

$$\hat{\mathbf{d}}_{i,p} = \hat{\mathbf{R}}_i \begin{pmatrix} (u - W/2)/\hat{f} \\ -(v - H/2)/\hat{f} \\ -1 \end{pmatrix}$$

1. L. Zirui Wang and Shangzhe Wu and Weidi Xie and Min Chen and Victor Adrian Prisacariu, “NeRF--: Neural Radiance Fields without Known Camera Parameters," arxiv 2021

NeRFmm https://github.com/ActiveVisionLab/nerfmm

清华大学车辆与运载学院
SCHOOL OF VEHICLE AND MOBILITY TSINGHUA UNIVERSITY

Scene	SSIM↑		LPIPS↓		PSNR↑		Camera Parameters Difference		
	colmap	ours	colmap	ours	colmap	ours	Δrot (deg)	Δtran	Δfocal (pixel)
Fern	0.64	0.61	0.47	0.50	22.22	21.67	1.78	0.029	153.5
Flower	0.71	0.71	0.36	0.37	25.25	25.34	4.84	0.016	13.2
Fortress	0.73	0.63	0.38	0.49	27.60	26.20	1.36	0.025	144.1
Horns	0.68	0.61	0.44	0.50	24.25	22.53	5.55	0.044	156.2
Leaves	0.52	0.53	0.47	0.47	18.81	18.88	3.90	0.016	59.0
Orchids	0.51	0.39	0.46	0.55	19.09	16.73	4.96	0.051	199.3
Room	0.87	0.84	0.40	0.44	27.77	25.84	2.77	0.030	331.8
Trex	0.74	0.72	0.41	0.44	23.19	22.67	4.67	0.036	89.3
Mean	0.68	0.63	0.42	0.47	23.52	22.48	3.73	0.031	143.3

LLFF数据集

Method	Camera Parameters Error (vs. GT)			NVS Quality	
	Δfocal (pixel)	Δrot (deg)	Δtran	SSIM↑	PSNR↑
colmap	14.89	13.65	0.0127	0.90	33.92
ours	20.55	4.45	0.0654	0.90	33.24

BLEFF数据集（本文提出的）

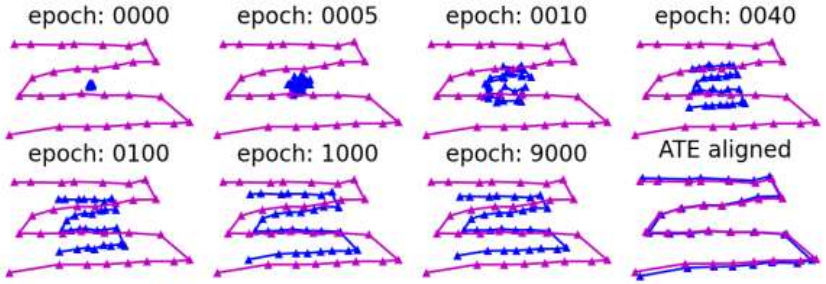


Figure 7: Pose optimisation during training, visualised on xy-plane (purple: COLMAP, blue: ours).

1. L. Zirui Wang and Shangzhe Wu and Weidi Xie and Min Chen and Victor Adrian Prisacariu, “NeRF--: Neural Radiance Fields without Known Camera Parameters," arxiv 2021

NeRFmm https://github.com/ActiveVisionLab/nerfmm

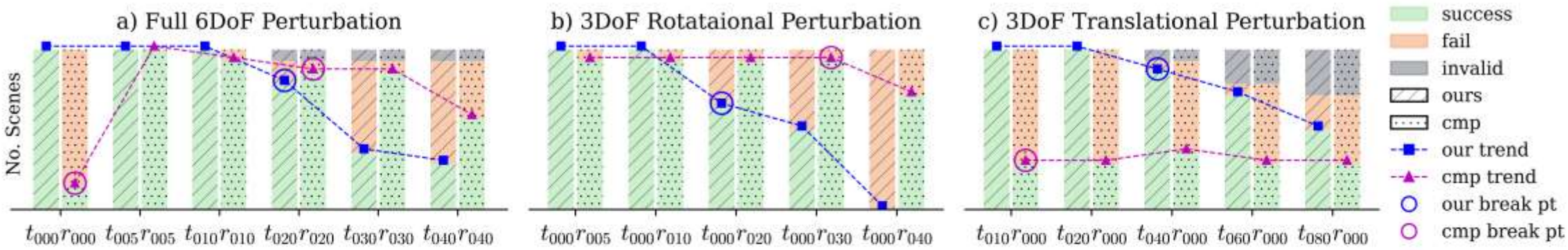


Figure 8: Breaking point analysis for camera parameter estimation in three groups of perturbation experiments: a) full 6DoF, b) 3DoF rotation, and c) 3DoF translation. COLMAP is shorten to "cmp" in the graph. We define a breaking point to the trajectory variant where a method starts failing in an experiment group. In group a) with full 6DoF noise, both methods start failing at $t_{020}r_{020}$. In 3DoF perturbations, our method performs more stable in translation perturbations but less stable in rotation perturbations. Note that COLMAP also faces degenerate issues with $t_{000}r_{000}$. Check Section 6.3.2 for more details on degenerate cases.

1. L. Zirui Wang and Shangzhe Wu and Weidi Xie and Min Chen and Victor Adrian Prisacariu, “NeRF--: Neural Radiance Fields without Known Camera Parameters," arxiv 2021

iMAP

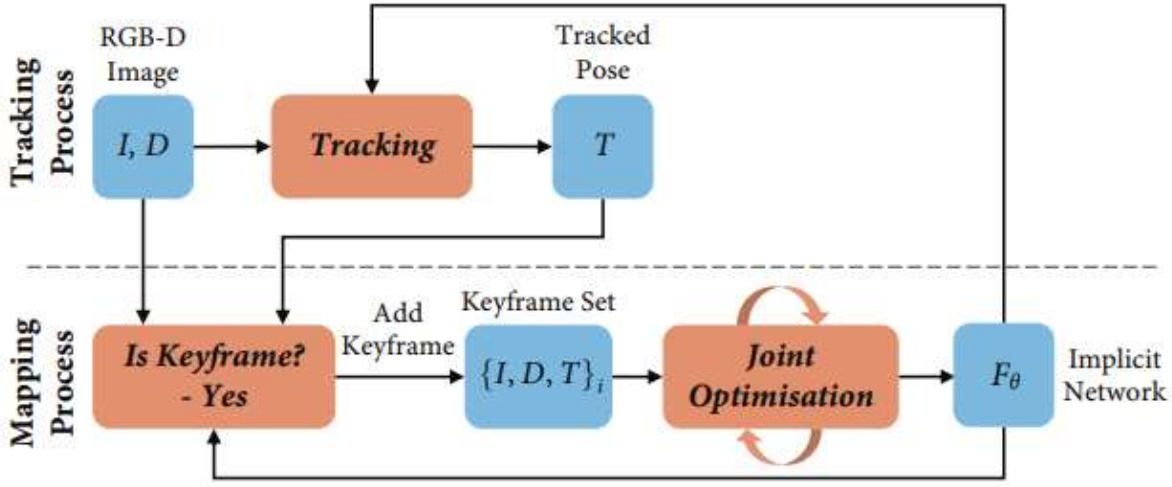


Figure 2: iMAP system pipeline.

1. E. Sucar, S. Liu, J. Ortiz and A. J. Davison, "iMAP: Implicit Mapping and Positioning in Real-Time," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6209-6218, doi: 10.1109/ICCV48922.2021.00617.

3.2. Implicit Scene Neural Network

Following the network architecture in NeRF [15], we use an MLP with 4 hidden layers of feature size 256, and two output heads that map a 3D coordinate $\mathbf{p} = (x, y, z)$ to a colour and volume density value: $F_\theta(\mathbf{p}) = (\mathbf{c}, \rho)$. Unlike NeRF, we do not take into account viewing directions as we are not interested in modelling specularities.

We apply the Gaussian positional embedding proposed in Fourier Feature Networks [32] to lift the input 3D coordinate into n-dimensional space: $\sin(\mathbf{Bp})$, with $\mathbf{B}$ an $[n \times 3]$ matrix sampled from a normal distribution with standard deviation σ. This embedding serves as input to the MLP and is also concatenated to the second activation layer of the network. Taking inspiration from SIREN [27], we allow optimisation of the embedding matrix $\mathbf{B}$, implemented as a single fully-connected layer with sine activation.

3.3. Depth and Colour Rendering

Our new differentiable rendering engine, inspired by NeRF [15] and NodeSLAM [31], queries the scene network to obtain depth and colour images from a given view.

Given a camera pose T_{WC} and a pixel coordinate $[u, v]$, we first back-project a normalised viewing direction and transform it into world coordinates: $\mathbf{r} = T_{WC} K^{-1} [u, v]$, with the camera intrinsics matrix K . We take a set of N samples along the ray $\mathbf{p}_i = d_i \mathbf{r}$ with corresponding depth values $\{d_1, \cdots, d_N\}$, and query the network for a colour and volume density $(\mathbf{c}_i, \rho_i) = F_\theta(\mathbf{p}_i)$. We follow the strati-

给定相机位姿，渲染出对应的图像$\hat{I}$和深度图$\hat{D}$

1. E. Sucar, S. Liu, J. Ortiz and A. J. Davison, "iMAP: Implicit Mapping and Positioning in Real-Time," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6209-6218, doi: 10.1109/ICCV48922.2021.00617.

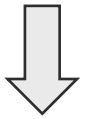
iMAP

Joint optimization

给定关键帧及对应的RGB-D观测，度量渲染结果与真实观测间的颜色和深度差异：

$$L_p = \frac{1}{M}\sum_{i=1}^{W}\sum_{(u,v)\in s_i} e_i^p[u,v].$$

$$L_g = \frac{1}{M}\sum_{i=1}^{W}\sum_{(u,v)\in s_i} \frac{e_i^g[u,v]}{\sqrt{\hat{D}_{var}[u,v]}}.$$



$$\min_{\theta,\{T_i\}} \left(L_g + \lambda_p L_p\right).$$

Camera Tracking

固定NeRF参数，使用与joint optimization完全一样的loss function，但只优化最后一帧的位姿

Tracking线程是并行运行的，运算效率更高

跟踪得到的位姿初值会在mapping阶段进行优化

1. E. Sucar, S. Liu, J. Ortiz and A. J. Davison, "iMAP: Implicit Mapping and Positioning in Real-Time," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6209-6218, doi: 10.1109/ICCV48922.2021.00617.

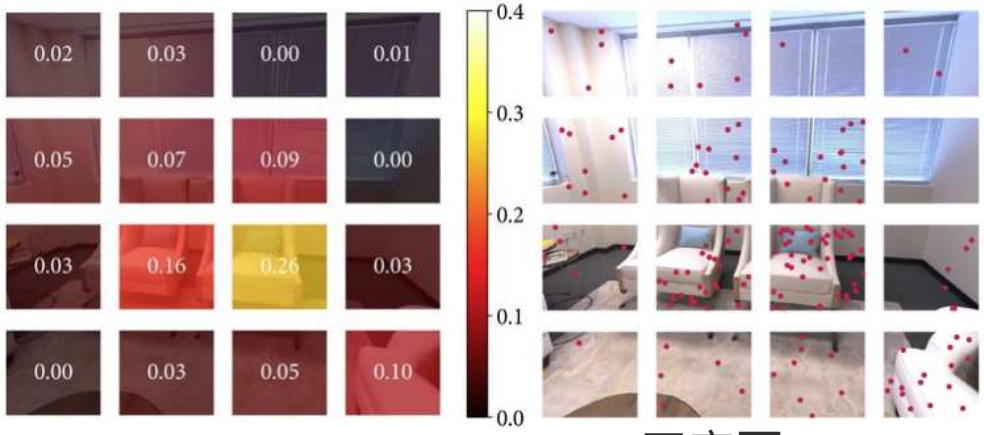


Image active sampling示意图

Image Active Sampling

Each joint optimisation iteration is divided into two stages. First, we sample a set s_i of pixels, uniformly distributed across each of the keyframe's depth and colour images. These pixels are used to update the network and camera poses, and to calculate the loss statistics. For this, we divide each image into an $[8\times8]$ grid, and calculate the average loss inside each square region R_j, $j = \{1, 2, \cdots, 64\}$:

$$L_i[j] = \frac{1}{|r_j|} \sum_{(u,v)\in r_j} e_i^g[u,v] + e_i^p[u,v], \tag{7}$$

where $r_j = s_i \cap R_j$ are pixels uniformly sampled from R_j. We normalise these statistics into a probability distribution:

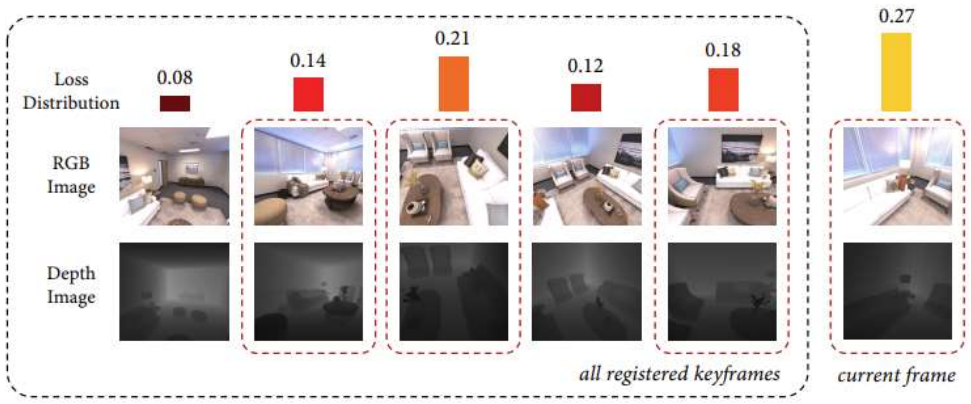
$$f_i[j] = \frac{L_i[j]}{\sum_{m=1}^{64} L_i[m]}. \tag{8}$$

We use this distribution to re-sample a new set of $n_i \cdot f_i[j]$ uniform samples per region (n_i is the total samples in each keyframe), allocating more samples to regions with high loss. The scene network is updated with the loss from active samples (in camera tracking only uniform sampling is used). Image active sampling is illustrated in Fig. 3.

1. E. Sucar, S. Liu, J. Ortiz and A. J. Davison, "iMAP: Implicit Mapping and Positioning in Real-Time," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6209-6218, doi: 10.1109/ICCV48922.2021.00617.

Keyframe Selection

$$P = \frac{1}{|s|} \sum_{(u,v)\in s} \mathbb{1}\left(\frac{\left| D[u,v] - \hat{D}[u,v] \right|}{D[u,v]} < t_D \right).$$



Keyframe active sampling示意图

Active and Bounded Keyframe Selection

Keyframe Active Sampling In iMAP, we continuously optimise our scene map with a set of selected keyframes, serving as a memory bank to avoid network forgetting. We wish to allocate more samples to keyframes with a higher loss, because they relate to regions which are newly explored, highly detailed, or that the network started to forget. We follow a process analogous to image active sampling and allocate n_i samples to each keyframe, proportional to the loss distribution across keyframes, See Fig. 4.

Bounded Keyframe Selection Our keyframe set keeps growing as the camera moves to new and unexplored regions. To bound joint optimisation computation, we choose a fixed number (3 in the live system) of keyframes at each iteration, randomly sampled according to the loss distribution. We always include the last keyframe and the current live frame in joint optimisation, to compose a bounded window with $W = 5$ constantly changing frames. See Fig. 4.

1. E. Sucar, S. Liu, J. Ortiz and A. J. Davison, "iMAP: Implicit Mapping and Positioning in Real-Time," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6209-6218, doi: 10.1109/ICCV48922.2021.00617.

iMAP

Reconstruction

		room-0	room-1	room-2	office-0	office-1	office-2	office-3	office-4	Avg.
iMAP	**# Keyframes**	11	12	12	10	11	10	14	11	13.37
	Acc. [cm]	3.58	3.69	4.68	5.87	3.71	4.81	4.27	4.83	4.43
	Comp. [cm]	5.06	4.87	5.51	6.11	5.26	5.65	5.45	6.59	**5.56**
	Comp. Ratio [< 5cm %]	83.91	83.45	75.53	77.71	79.64	77.22	77.34	77.63	**79.06**
TSDF Fusion	**Acc.** [cm]	4.21	3.08	2.88	2.70	2.66	4.27	4.07	3.70	**3.45**
	Comp. [cm]	5.04	4.35	5.40	10.47	10.29	6.43	6.26	4.78	6.63
	Comp. Ratio [< 5cm %]	76.90	79.87	77.79	79.60	71.93	71.66	65.87	77.11	75.09

Table 1: Reconstruction results for 8 indoor Replica scenes. We report the highest reached completion ratio in each scene along with the corresponding accuracy and completion values at that point.

Figure 11: Hole filling capacity of iMAP (top) against BAD-SLAM (bottom).

Localization

	fr1/desk (cm)	fr2/xyz (cm)	fr3/office (cm)
iMAP	4.9	2.0	5.8
BAD-SLAM	1.7	1.1	1.73
Kintinuous	3.7	2.9	3.0
ORB-SLAM2	1.6	0.4	1.0

Table 3: ATE RMSE in cm on TUM RGB-D dataset.

1. E. Sucar, S. Liu, J. Ortiz and A. J. Davison, "iMAP: Implicit Mapping and Positioning in Real-Time," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6209-6218, doi: 10.1109/ICCV48922.2021.00617.

Some extensions to iMAP

1. L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola and T. -Y. Lin, "**iNeRF: Inverting Neural Radiance Fields for Pose Estimation**," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp. 1323-1330, doi: 10.1109/IROS51168.2021.9636708.
2. Z. Zhu et al., "**NICE-SLAM: Neural Implicit Scalable Encoding for SLAM**," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 12776-12786, doi: 10.1109/CVPR52688.2022.01245.
Code: https://github.com/cvg/nice-slam
3. Zhu, Zihan et al. "**NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM.**" *ArXiv* abs/2302.03594 (2023): n. pag.
4. Rosinol, Antoni et al. "**NeRF-SLAM: Real-Time Dense Monocular SLAM with Neural Radiance Fields.**" *ArXiv* abs/2210.13641 (2022): n. pag.
Code: https://github.com/ToniRV/NeRF-SLAM

LocNeRF利用粒子滤波算法来估计后验概率：

$$\mathbb{P}\left(\boldsymbol{X}_t \mid \mathcal{M}, \mathcal{I}_{1:t}, \mathcal{O}_{1:t}\right)$$

Predication Step

$$\boldsymbol{X}_t = \boldsymbol{X}_{t-1} \cdot \mathcal{O}_t \cdot \boldsymbol{X}_\epsilon \quad , \quad \boldsymbol{X}_\epsilon = \mathrm{Exp}\left(\boldsymbol{\delta}\right)$$

Update Step

$$\mathbb{P}\left(\mathcal{I}_t \mid \boldsymbol{X}_t^i, \mathcal{M}\right) \Leftarrow w_t^i = \left(\frac{M}{\sum_{j=1}^{M}(\mathcal{I}_t(\boldsymbol{p}_j) - C(\boldsymbol{r}(\boldsymbol{p}_j, \boldsymbol{X}_t^i)))^2}\right)^4$$

Resampling Step

After the update step, we resample n particles from the set S_t with replacement, where each particle is sampled with probability w_t^i. As prescribed by standard particle filtering, the resampling step allows retaining particles that are more likely to correspond to good pose estimates while discarding less likely hypotheses.

1. Maggio, Dominic et al. “Loc-NeRF: Monte Carlo Localization using Neural Radiance Fields.” 2023 IEEE International Conference on Robotics and Automation (ICRA) (2022): 4018-4025.

LocNeRF

https://github.com/MIT-SPARK/Loc-NeRF

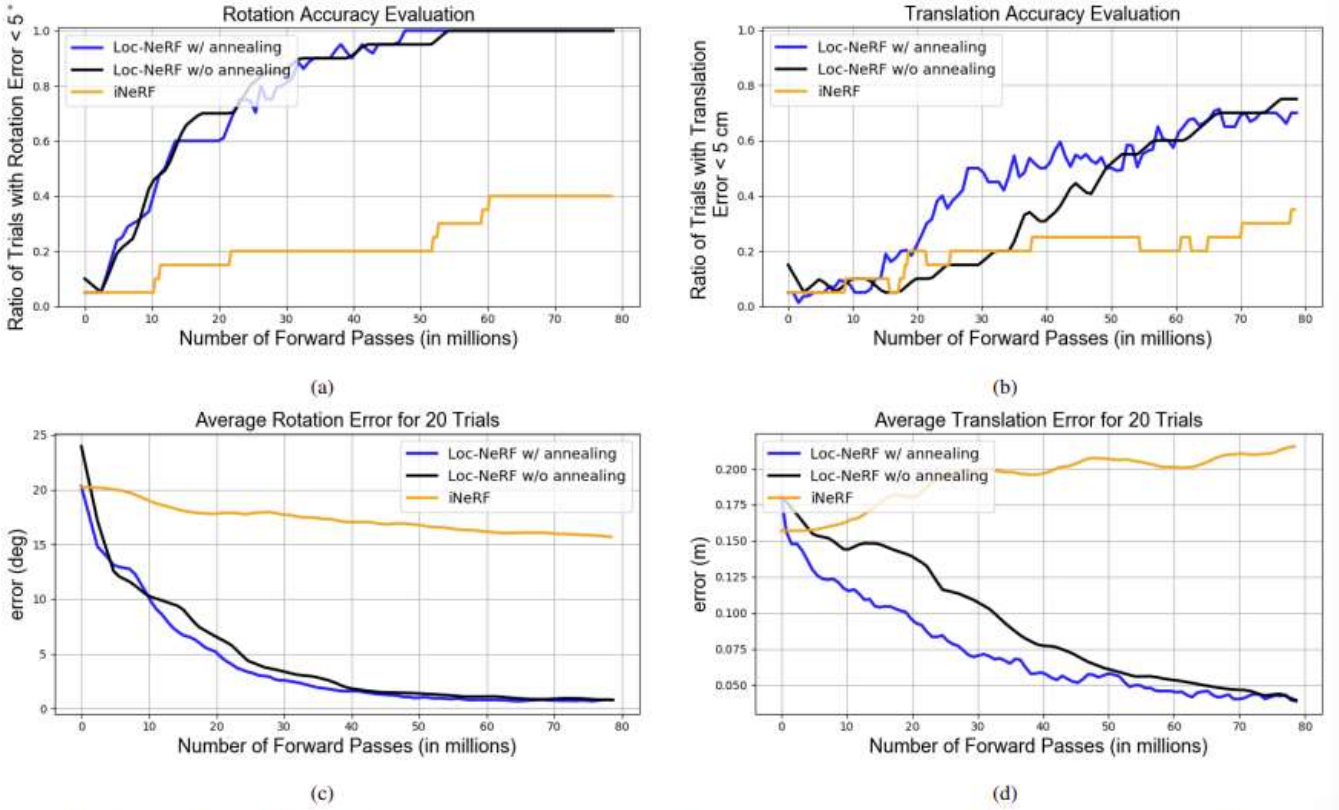


Fig. 2. Evaluation of Loc-NeRF and iNeRF on 20 camera poses from the LLFF dataset. As an ablation study of our annealing step, we also include results of Loc-NeRF without using Algorithm 1. (a) Ratio of trials with rotation error $< 5°$. (b) Ratio of trials with translation error < 5 cm. (c) Average rotation error. (d) Average translation error.

1. Maggio, Dominic et al. “Loc-NeRF: Monte Carlo Localization using Neural Radiance Fields.” 2023 IEEE International Conference on Robotics and Automation (ICRA) (2022): 4018-4025.

LocoNeRF

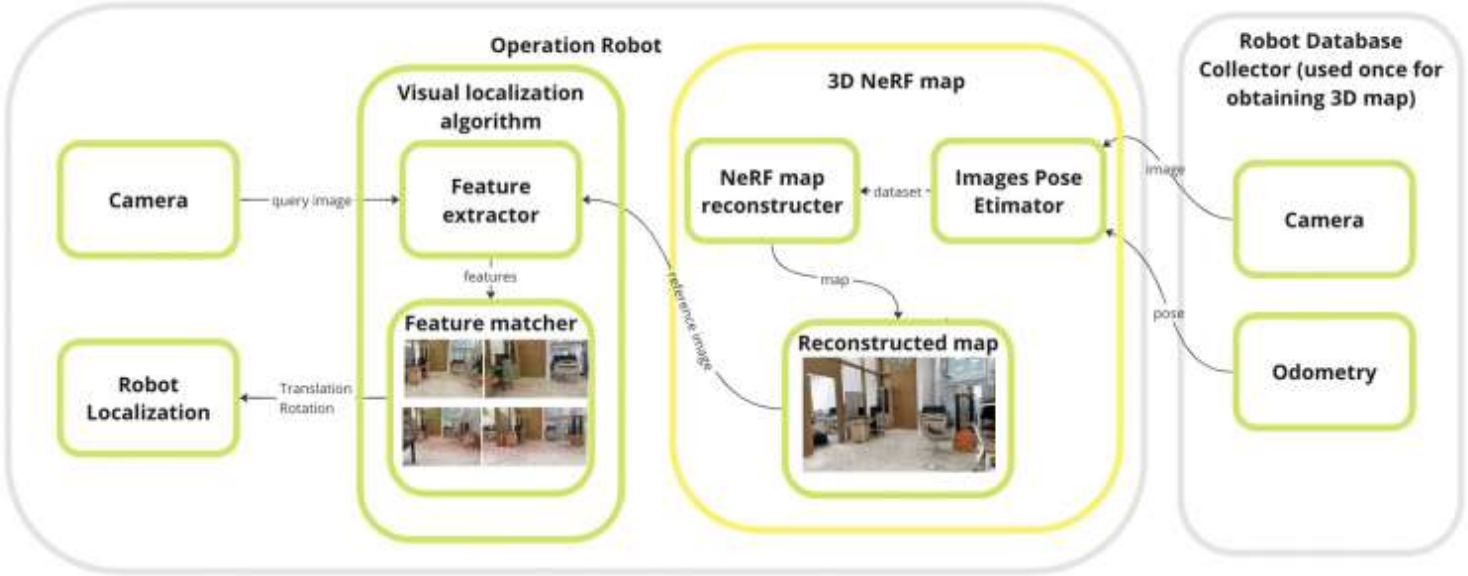
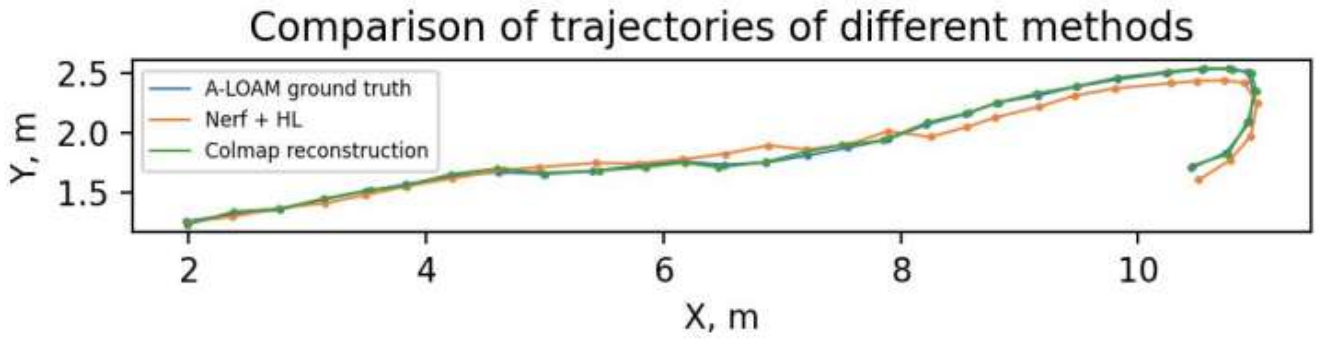


Fig. 4: Proposed visual localization based on Structure From Motion with Neural Radiance Fields map pipeline.



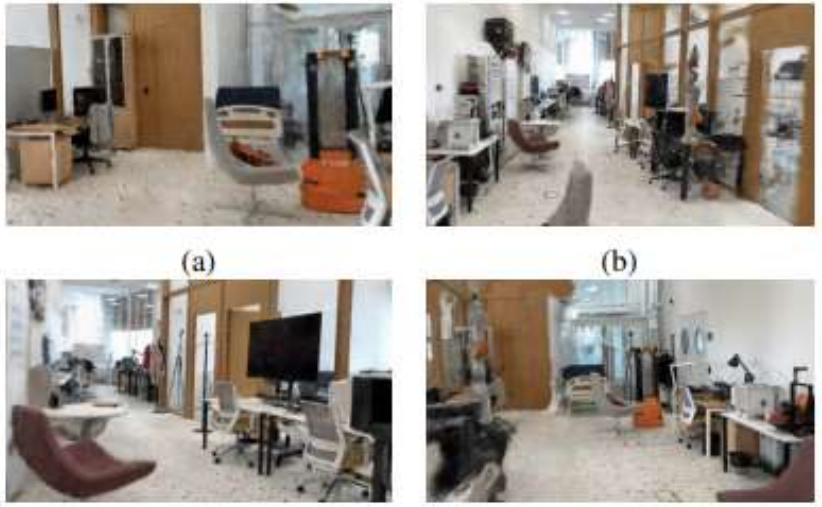


Fig. 7: The images rendered from 3D reconstruction based on nerfacto method

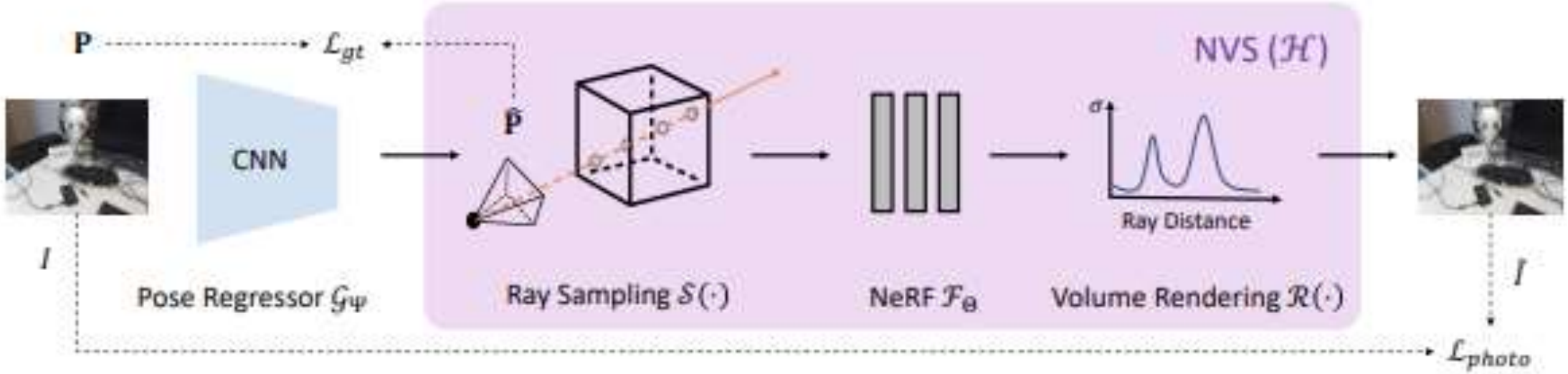
TABLE II: Localization experiment result metrics

Method	Trajectory Error, m	Rotation Error, rad	Map size, megabytes
COLMAP	0.022	0.012	400
SFM with NeRF	0.068	0.07	160

1. Nenashev, Artem et al. “LocoNeRF: A NeRF-based Approach for Local Structure from Motion for Precise Localization.” *ArXiv* abs/2310.05134 (2023): n. pag.

DPN

https://github.com/ActiveVisionLab/direct-posenet



$$\mathcal{L}_{photo}(\hat{I}, I) = \|\hat{I} - I\|_2.$$

$$\mathcal{L}_{gt} = \|\mathbf{P} - \hat{\mathbf{P}}\|_2,$$

$$\Psi^* = \arg\min_{\Psi}(\lambda_1 \mathcal{L}_{photo} + \lambda_2 \mathcal{L}_{gt}),$$

1. S. Chen, Z. Wang and V. Prisacariu, "Direct-PoseNet: Absolute Pose Regression with Photometric Consistency," 2021 International Conference on 3D Vision (3DV), London, United Kingdom, 2021, pp. 1175-1185, doi: 10.1109/3DV53792.2021.00125.

	without unlabeled data									with unlabeled data		
Scene	PN [22]	PN learned weights [21]	geo. PN [21]	LSTM PN [59]	Hourglass PN [30]	BranchNet [61]	DSO [10]	MapNet [6]	Direct-PN	MapNet+ [6]	MapNet+ PGO [6]	Direct-PN+U
Chess	0.32/8.12	0.14/4.50	0.13/4.48	0.24/5.77	0.15/6.17	0.18/5.17	0.17/8.13	0.08/3.25	0.10/3.52	0.10/3.17	0.09/3.24	0.09/2.77
Fire	0.47/14.4	0.27/11.8	0.27/11.3	0.34/11.9	0.27/10.8	0.34/8.99	0.19/65.0	0.27/11.7	0.27/8.66	0.20/9.04	0.20/9.29	0.16/4.87
Heads	0.29/12.0	0.18/12.1	0.17/13.0	0.21/13.7	0.19/11.6	0.20/14.2	0.61/68.2	0.18/13.3	0.17/13.1	0.13/11.1	0.12/8.45	0.10/6.64
Office	0.48/7.68	0.20/5.77	0.19/5.55	0.30/8.08	0.21/8.48	0.30/7.05	1.51/16.8	0.17/5.15	0.16/5.96	0.18/5.38	0.19/5.42	0.17/5.04
Pumpkin	0.47/8.42	0.25/4.82	0.26/4.75	0.33/7.00	0.25/7.0	0.27/5.10	0.61/15.8	0.22/4.02	0.19/3.85	0.19/3.92	0.19/3.96	0.19/3.59
Kitchen	0.59/8.64	0.24/5.52	0.23/5.35	0.37/8.83	0.27/10.2	0.33/7.40	0.23/10.9	0.23/4.93	0.22/5.13	0.20/5.01	0.20/4.94	0.19/4.79
Stairs	0.47/13.8	0.37/10.6	0.35/12.4	0.40/13.7	0.29/12.5	0.38/10.3	0.26/21.3	0.30/12.1	0.32/10.61	0.30/13.4	0.27/10.6	0.24/8.52
Average	0.44/10.44	0.24/7.87	0.23/8.12	0.31/9.85	0.23/9.53	0.29/8.30	0.26/29.4	0.21/7.77	0.20/7.26	0.19/7.29	0.18/6.55	0.16/5.17

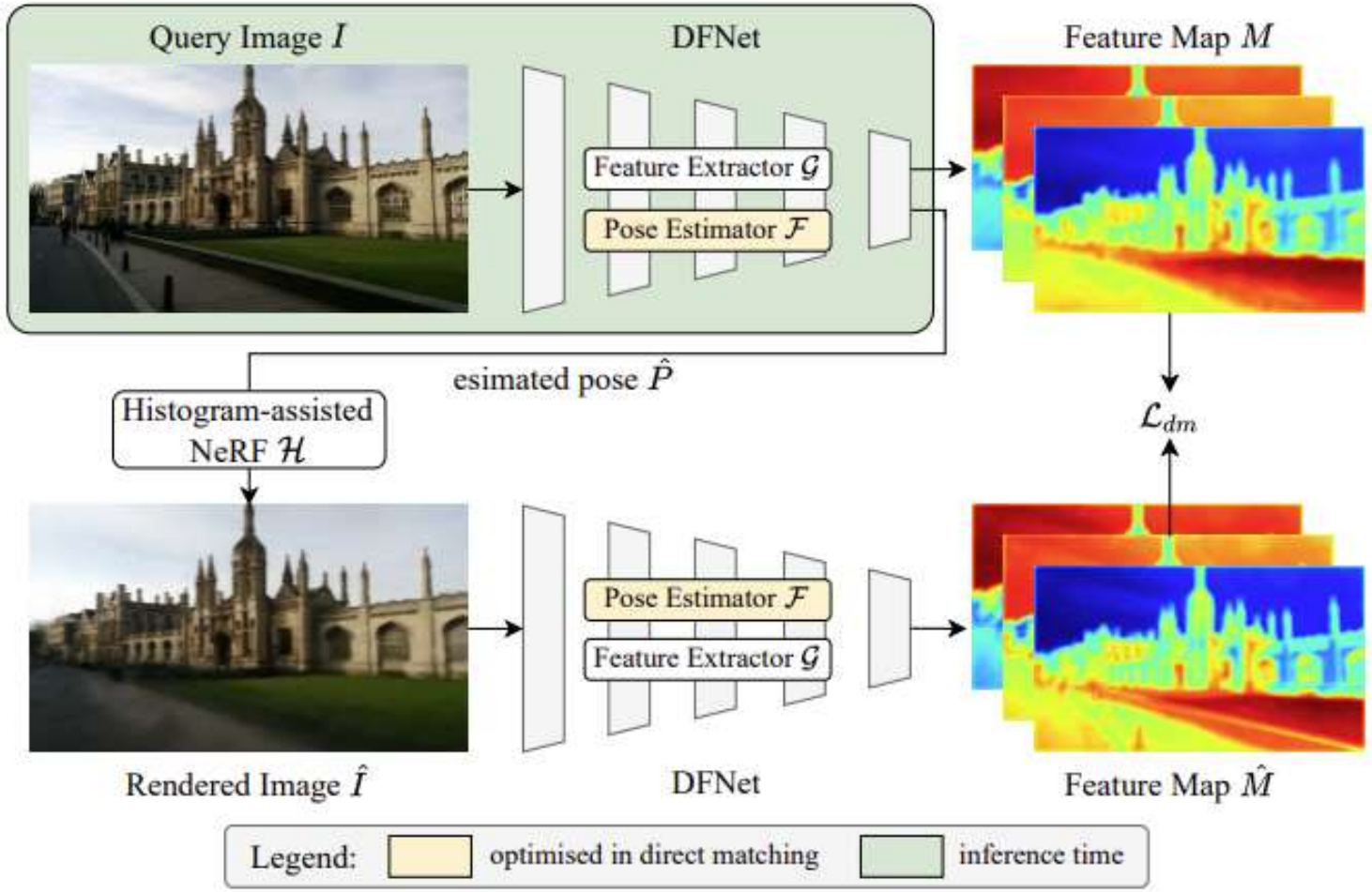
Table 1: Pose regression results on 7 Scenes datasets. We compare our method with both direct matching and absolute pose regression methods, in median translation error (m) and rotation error (°). Bottom row is the average of median errors of all scenes. PN denotes PoseNet. Numbers in red represent the best performance with or without unlabeled data.

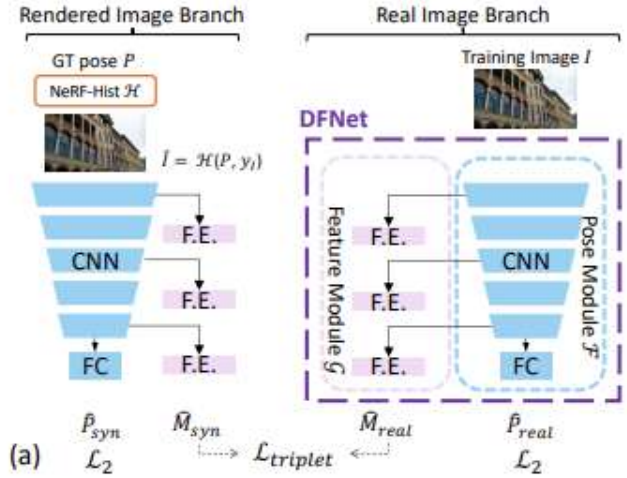
	without unlabeled data			with unlabeled data	
Model	PoseNet+logq [6]	MapNet [6]	Direct-PN	MapNet+PGO [6]	Direct-PN+U
Avg. Median	0.23m, 8.49°	0.21m, 7.77°	**0.20m, 7.26°**	0.18m, 6.55°	**0.16m, 5.17°**
Avg. Mean	0.28m, 10.43°	0.27m, 10.08°	**0.25m, 8.98°**	0.22m, 7.89°	**0.21m, 7.02°**

Table 7: A comparison of average median errors and average mean errors on the 7-Scenes dataset.

1. S. Chen, Z. Wang and V. Prisacariu, "Direct-PoseNet: Absolute Pose Regression with Photometric Consistency," 2021 International Conference on 3D Vision (3DV), London, United Kingdom, 2021, pp. 1175-1185, doi: 10.1109/3DV53792.2021.00125.

DFNet https://github.com/hughplay/DFNet





$$\mathcal{L}_{gt} = \|\mathbf{P} - \hat{\mathbf{P}}\|_2,$$

$$\mathcal{L}_{dm} = \sum_i (1 - \cos(m_i, \tilde{m}_i))$$

$$\cos(m_i, \tilde{m}_i) = \frac{m_i \cdot \tilde{m}_i}{\|m_i\|_2 \cdot \|\tilde{m}_i\|_2}.$$

1. Chen, S., Li, X., Wang, Z., Prisacariu, V.A. DFNet: Enhance Absolute Pose Regression with Direct Feature Matching. In Proceedings of European Conference on Computer Vision, 2022

Closing the Domain Gap

原本的triplet loss：

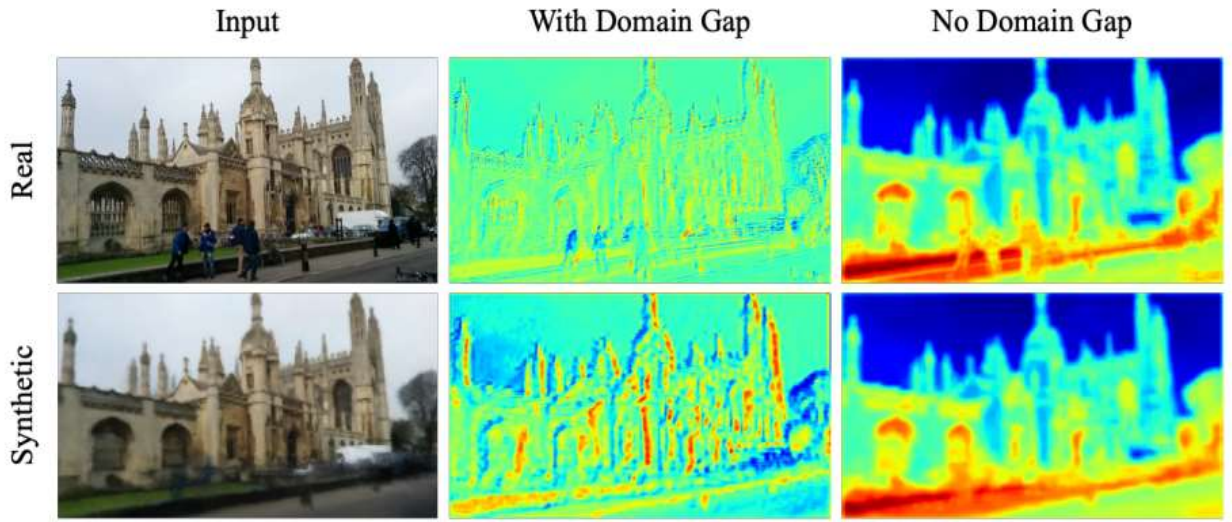
$$\mathcal{L}_{triplet}^{ori} = \max\left\{d(M_{real}^{P}, M_{syn}^{P}) - d(M_{real}^{P}, M_{syn}^{\bar{P}}) + \text{margin}, 0\right\}$$

改进的triplet loss：

$$\mathcal{L}_{triplet} = \max\left\{d(M_{real}^{P}, M_{syn}^{P}) - q_{\ominus} + \text{margin}, 0\right\}$$

$$q_{\ominus} = \min\left\{d(M_{real}^{P}, M_{real}^{\bar{P}}), d(M_{real}^{P}, M_{syn}^{\bar{P}}), d(M_{syn}^{P}, M_{real}^{\bar{P}}), d(M_{syn}^{P}, M_{syn}^{\bar{P}})\right\},$$

$$\mathcal{L}_{DFNet} = \mathcal{L}_{triplet} + \mathcal{L}_{RVS} + \frac{1}{2}(\|P - \hat{P}_{real}\|_2 + \|P - \hat{P}_{syn}\|_2)$$



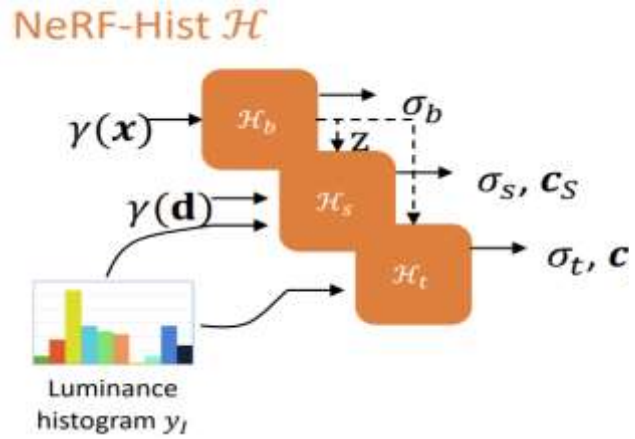
1. Chen, S., Li, X., Wang, Z., Prisacariu, V.A. DFNet: Enhance Absolute Pose Regression with Direct Feature Matching. In Proceedings of European Conference on Computer Vision, 2022

DFNet

https://github.com/hughplay/DFNet

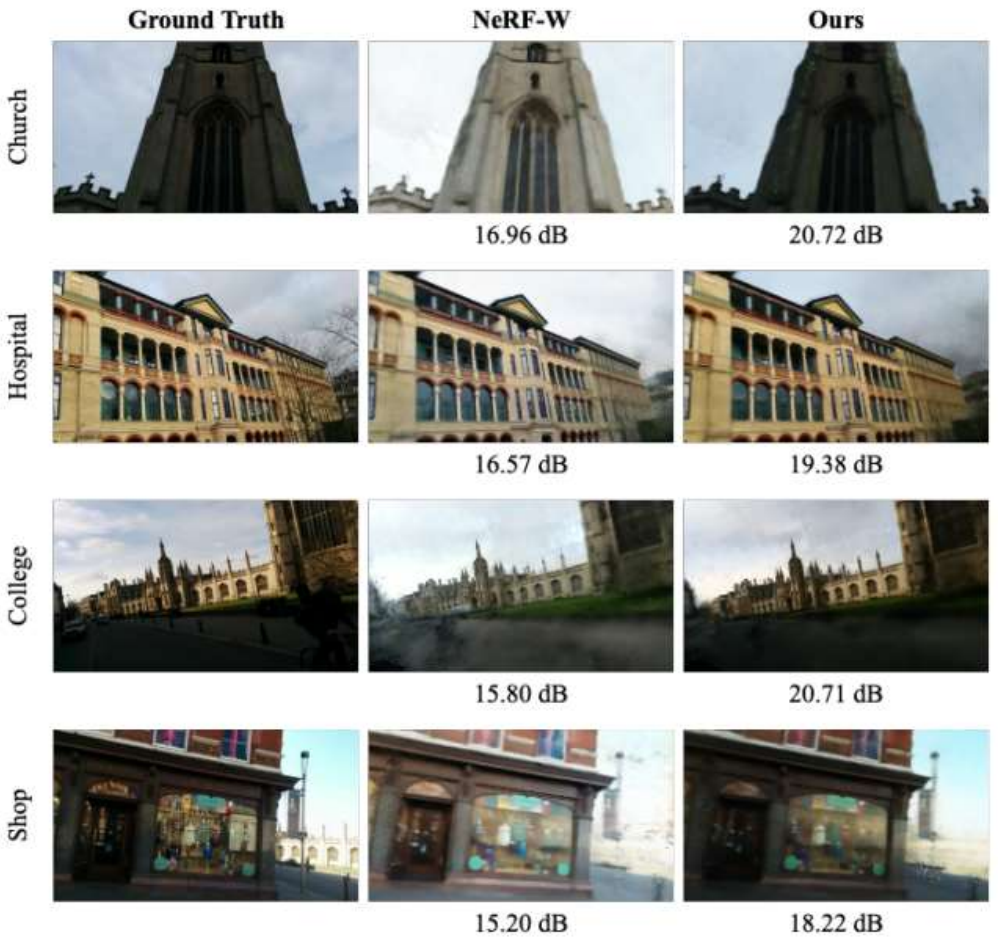
清华大学车辆与运载学院
SCHOOL OF VEHICLE AND MOBILITY TSINGHUA UNIVERSITY

Histogram-assisted NeRF



1. A base network $\mathcal{H}_b$ that provides a density estimation σ_b and a hidden state $\mathbf{z}$ for a coarse estimation: $[\sigma_b, \mathbf{z}] = \mathcal{H}_b(\gamma(\mathbf{x}))$.
2. A static network $\mathcal{H}_s$ to model density σ_s and radiance $\mathbf{c}_s$ for static structure and appearance: $[\sigma_s, \mathbf{c}_s] = \mathcal{H}_s(\mathbf{z}, \gamma(\mathbf{d}), \mathbf{y}_I)$.
3. A transient network $\mathcal{H}_t$ to model density σ_t, radiance $\mathbf{c}_t$ and an uncertainty estimation β for dynamic objects: $[\sigma_t, \mathbf{c}_t, \beta] = \mathcal{H}_t(\mathbf{z}, \mathbf{y}_I)$.

Both the static and the transient network are conditioned on a histogram-based embedding $\mathbf{y}_I \in \mathbb{R}^{C_y}$, which is mapped from a N_b bins histogram. The histogram is computed on the luma channel Y of a target image in YUV space. We found this approach works well in a direct matching context, not only in feature-metric space but also in photometric space.



1. Chen, S., Li, X., Wang, Z., Prisacariu, V.A. DFNet: Enhance Absolute Pose Regression with Direct Feature Matching. In Proceedings of European Conference on Computer Vision, 2022

DFNet https://github.com/hughplay/DFNet

清华大学车辆与运载学院
SCHOOL OF VEHICLE AND MOBILITY TSINGHUA UNIVERSITY

Table 1. Pose regression results on 7-Scenes dataset. We compare DFNet and DFNet$_{dm}$ (DFNet with feature-metric direct matching) with prior single-frame APR methods and unlabeled training methods, in median translation error (m) and rotation error (°). Note that MapNet+ and MapNet+PGO are sequential methods with unlabeled training. Numbers in **bold** represent the best performance.

	Methods	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Average
1-frame APR	PoseNet(PN)[14]	0.32/8.12	0.47/14.4	0.29/12.0	0.48/7.68	0.47/8.42	0.59/8.64	0.47/13.8	0.44/10.4
	PN Learn σ^2[13]	0.14/4.50	0.27/11.8	0.18/12.1	0.20/5.77	0.25/4.82	0.24/5.52	0.37/10.6	0.24/7.87
	geo. PN[13]	0.13/4.48	0.27/11.3	0.17/13.0	0.19/5.55	0.26/4.75	0.23/5.35	0.35/12.4	0.23/8.12
	LSTM PN[36]	0.24/5.77	0.34/11.9	0.21/13.7	0.30/8.08	0.33/7.00	0.37/8.83	0.40/13.7	0.31/9.85
	Hourglass PN[19]	0.15/6.17	0.27/10.8	0.19/11.6	0.21/8.48	0.25/7.0	0.27/10.2	0.29/12.5	0.23/9.53
	BranchNet[38]	0.18/5.17	0.34/8.99	0.20/14.2	0.30/7.05	0.27/5.10	0.33/7.40	0.38/10.3	0.29/8.30
	MapNet[4]	0.08/3.25	0.27/11.7	0.18/13.3	0.17/5.15	0.22/4.02	0.23/**4.93**	0.30/12.1	0.21/7.77
	Direct-PN[5]	0.10/3.52	0.27/8.66	0.17/13.1	0.16/5.96	0.19/3.85	0.22/5.13	0.32/10.6	0.20/7.26
	TransPoseNet[29]	0.08/5.68	0.24/10.6	0.13/12.7	0.17/6.34	0.17/5.6	0.19/6.75	0.30/7.02	0.18/7.78
	MS-Transformer[28]	0.11/4.66	0.24/9.60	0.14/12.2	0.17/5.66	0.18/4.44	**0.17**/5.94	0.17/5.94	0.18/7.28
	DFNet (ours)	**0.05/1.88**	**0.17/6.45**	**0.06/3.63**	**0.08/2.48**	**0.10/2.78**	0.22/5.45	**0.16/3.29**	**0.12/3.71**
Unlabel Data	MapNet$_{+(seq.)}$[4]	0.10/3.17	0.20/9.04	0.13/11.1	0.18/5.38	0.19/3.92	0.20/5.01	0.30/13.4	0.19/7.29
	MapNet$_{+PGO(seq.)}$[4]	0.09/3.24	0.20/9.29	0.12/8.45	0.19/5.42	0.19/3.96	0.20/4.94	0.27/10.6	0.18/6.55
	Direct-PN+U[5]	0.09/2.77	0.16/4.87	0.10/6.64	0.17/5.04	0.19/3.59	0.19/4.79	0.24/8.52	0.16/5.17
	DFNet$_{dm}$ (ours)	**0.04/1.48**	**0.04/2.16**	**0.03/1.82**	**0.07/2.01**	**0.09/2.26**	**0.09/2.42**	**0.14/3.31**	**0.07/2.21**

Table 2. Single-frame APR results on Cambridge dataset. We report the median position and orientation errors in $m/°$ and the respective rankings over scene average as in [29,28]. The best results is highlighted in **bold**. For fair comparisons, we omit prior APR methods which did not publish results in Cambridge.

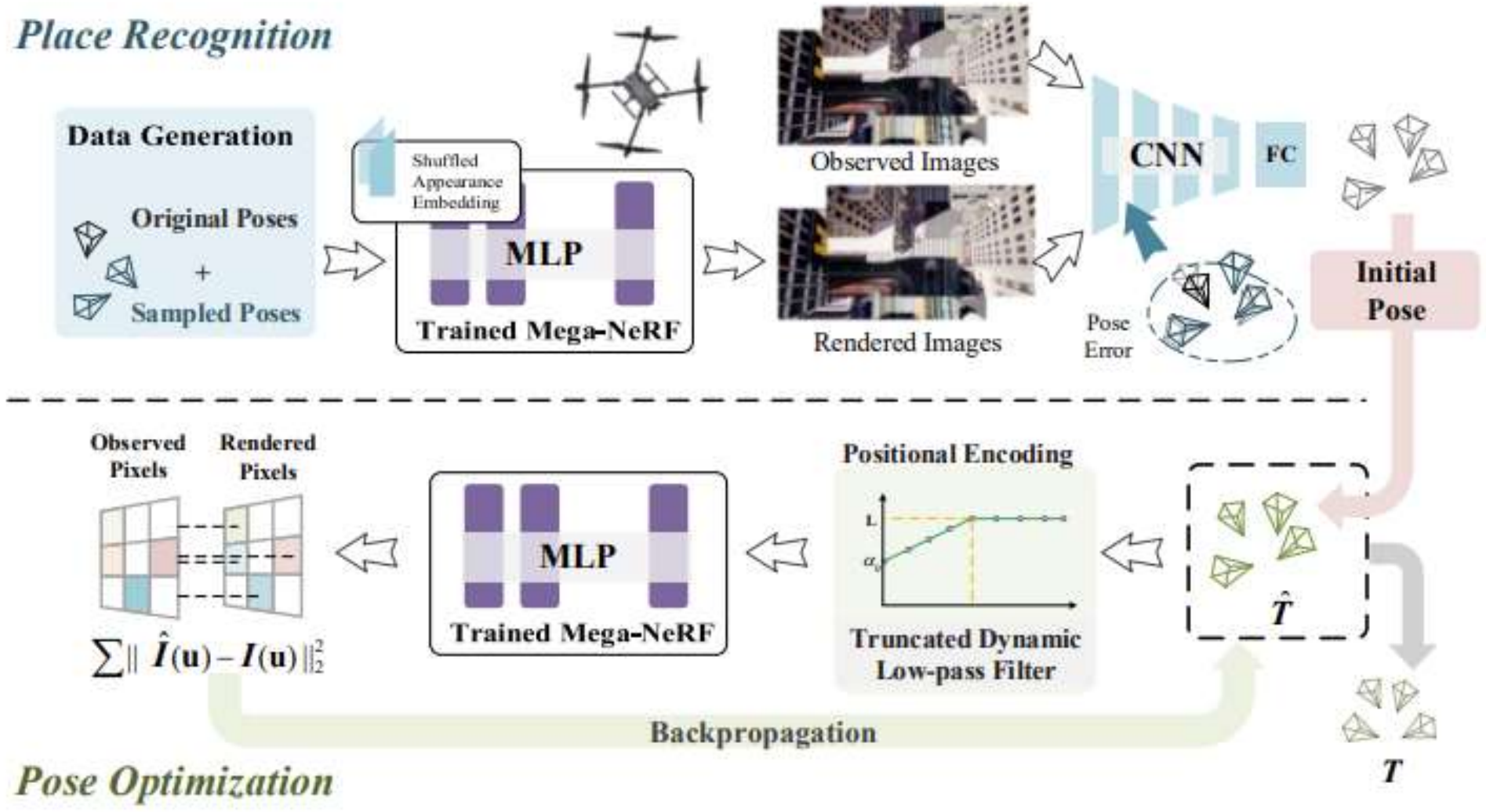
Methods	Kings	Hospital	Shop	Church	Average	Ranks	Final Rank
PoseNet(PN)[14]	1.66/4.86	2.62/4.90	1.41/7.18	2.45/7.96	2.04/6.23	9/9	9
PN Learn σ^2[13]	0.99/1.06	2.17/2.94	1.05/3.97	1.49/3.43	1.43/2.85	6/3	5
geo. PN[13]	0.88/1.04	3.20/3.29	0.88/3.78	1.57/3.32	1.63/2.86	7/4	6
LSTM PN[36]	0.99/3.65	1.51/4.29	1.18/7.44	1.52/6.68	1.30/5.51	5/8	7
MapNet[4]	1.07/1.89	1.94/3.91	1.49/4.22	2.00/4.53	1.63/3.64	7/7	8
TransPoseNet[29]	0.60/2.43	1.45/3.08	0.55/3.49	1.09/4.94	0.91/3.50	2/6	3
MS-Transformer[28]	0.83/1.47	1.81/2.39	0.86/3.07	1.62/3.99	1.28/2.73	4/2	2
DFNet (ours)	0.73/2.37	2.00/2.98	0.67/2.21	1.37/4.03	1.19/2.90	3/5	3
DFNet$_{dm}$ (ours)	**0.43/0.87**	**0.46/0.87**	**0.16/0.59**	**0.50/1.49**	**0.39/0.96**	**1/1**	**1**

Table 3. Comparison between our method and sequential-based APR methods and 3D structure-based methods.

	3D	Seq. APR				1-frame
Methods	AS[27]	MapNet +PGO[4]	CoordiNet[21]	CoordiNet +Lens[22]	VLocNet[34]	DFNet$_{dm}$
Chess	0.04/2.0	0.09/3.24	0.14/6.7	0.03/1.3	0.04/1.71	0.04/1.48
Fire	0.03/1.5	0.20/9.29	0.27/11.6	0.10/3.7	0.04/5.34	0.04/2.16
Heads	0.02/1.5	0.12/8.45	0.13/13.6	0.07/5.8	0.05/6.65	0.03/1.82
Office	0.09/3.6	0.19/5.42	0.21/8.6	0.07/1.9	0.04/1.95	0.07/2.01
Pumpkin	0.08/3.1	0.19/3.96	0.25/7.2	0.08/2.2	0.04/2.28	0.09/2.26
Kitchen	0.07/3.4	0.20/4.94	0.26/7.5	0.09/2.2	0.04/2.21	0.09/2.42
Stairs	0.03/2.2	0.27/10.6	0.28/12.9	0.14/3.6	0.10/6.48	0.14/3.31
Average	0.05/2.5	0.18/6.55	0.22/9.7	0.08/3.0	**0.05**/3.80	0.07/**2.21**
Kings	0.42/0.6	-	0.70/2.92	0.33/0.5	0.84/1.42	0.43/0.87
Hospital	0.44/1.0	-	0.97/2.08	0.44/0.9	1.08/2.41	0.46/0.87
Shop	0.12/0.4	-	0.73/4.69	0.27/1.6	0.59/3.53	0.16/0.59
Church	0.19/0.5	-	1.32/3.56	0.53/1.6	0.63/3.91	0.50/1.49
Average	0.29/0.63	-	0.92/2.58	**0.39**/1.15	0.78/2.82	**0.39/0.96**

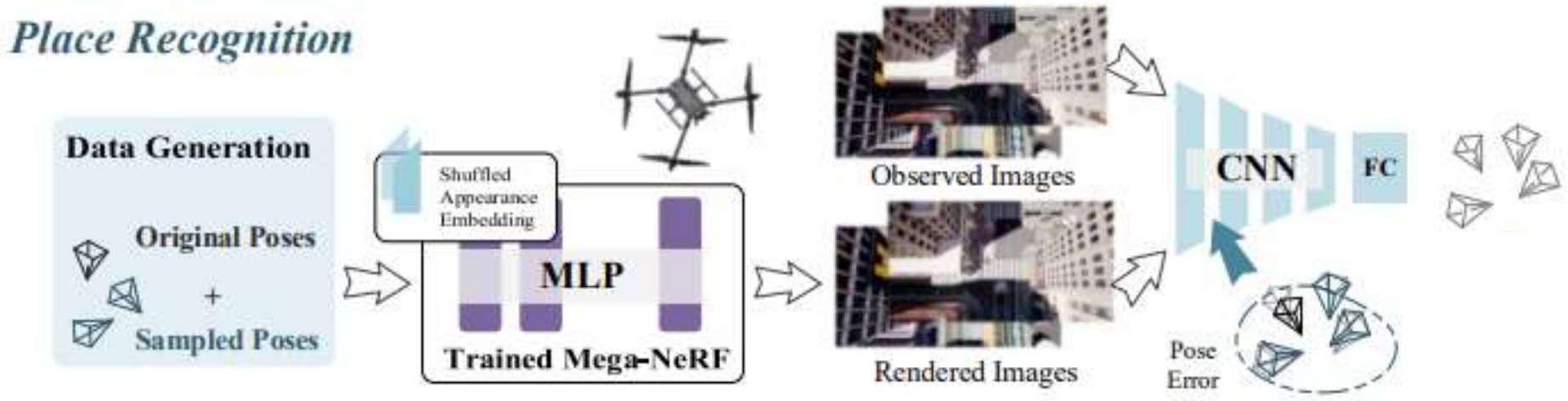
1. Chen, S., Li, X., Wang, Z., Prisacariu, V.A. DFNet: Enhance Absolute Pose Regression with Direct Feature Matching. In Proceedings of European Conference on Computer Vision, 2022

LATITUDE https://github.com/jike5/LATITUDE



1. Z. Zhu et al., "LATITUDE: Robotic Global Localization with Truncated Dynamic Low-pass Filter in City-scale NeRF," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 8326-8332, doi: 10.1109/ICRA48891.2023.10161570.

LATITUDE https://github.com/jike5/LATITUDE



To begin, we uniformly sample several positions in a horizontal $H \times W$ rectangle area around each position in training set where H and W is a given parameter. Then, we need to define the camera orientation attached to these positions for each virtual camera pose. To avoid degenerate views, we copy the pose orientations of the training set, add random perturbations on each axis drawn evenly in $[-\theta, \theta]$, where θ is the maximum amplitude of the perturbation.

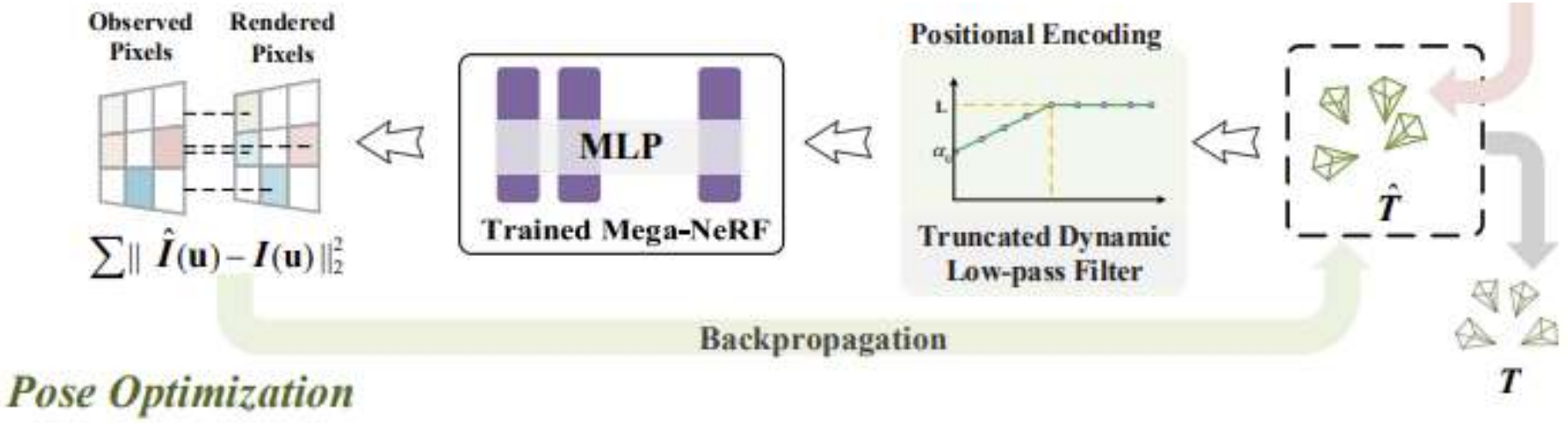
$$\mathcal{L}_{real} = \|\hat{\mathbf{x}}_{real} - \mathbf{x}_{real}\|_2 + \gamma \left\| \hat{\mathbf{q}}_{real} - \frac{\mathbf{q}_{real}}{\|\mathbf{q}_{real}\|} \right\|_2$$

$$\mathcal{L}_{syn} = \|\hat{\mathbf{x}}_{syn} - \mathbf{x}_{syn}\|_2 + \gamma \left\| \hat{\mathbf{q}}_{syn} - \frac{\mathbf{q}_{syn}}{\|\mathbf{q}_{syn}\|} \right\|_2$$

$$\mathcal{L} = \mathcal{L}_{real} + \beta \mathcal{L}_{syn}$$

1. Z. Zhu et al., "LATITUDE: Robotic Global Localization with Truncated Dynamic Low-pass Filter in City-scale NeRF," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 8326-8332, doi: 10.1109/ICRA48891.2023.10161570.

LATITUDE https://github.com/jike5/LATITUDE



$$\hat{\xi} = \arg\min_{\xi \in \mathfrak{se}(3)} \mathcal{L}(\xi | \hat{T}_0, I, \theta)$$

$$\hat{T} = \exp(\hat{\xi}) \hat{T}_0$$

$$\Longrightarrow \mathcal{L} = \sum_{\mathbf{u} \in \mathcal{R}} \| \mathcal{I}\left(\mathcal{F}(TK\mathbf{u}; \theta)\right) - I(\mathbf{u}) \|_2^2$$

1. Z. Zhu et al., "LATITUDE: Robotic Global Localization with Truncated Dynamic Low-pass Filter in City-scale NeRF," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 8326-8332, doi: 10.1109/ICRA48891.2023.10161570.

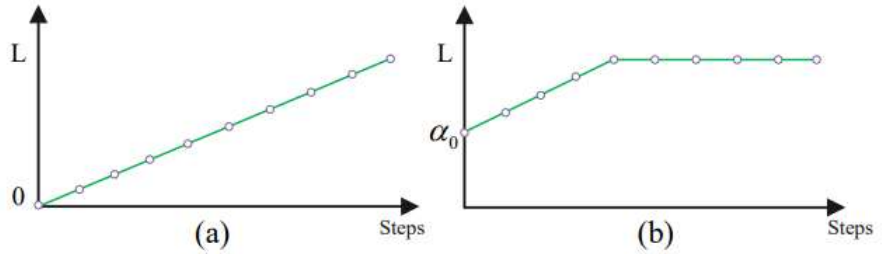
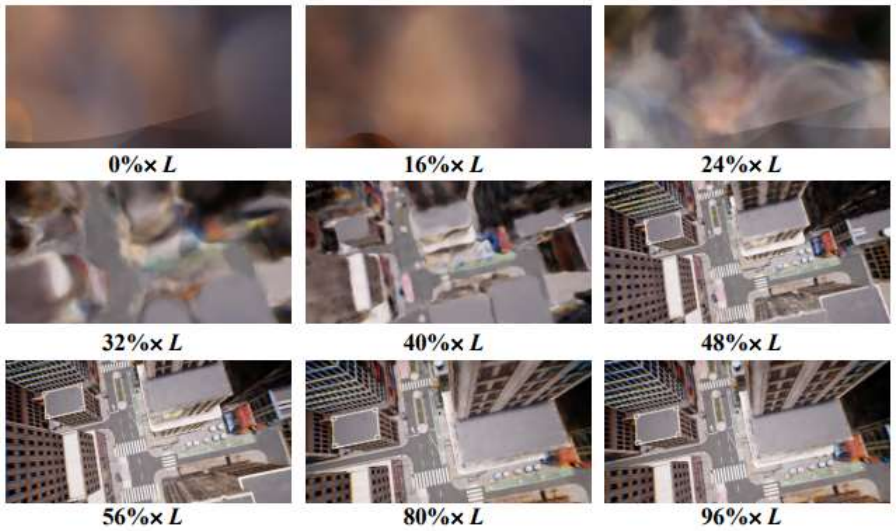


Fig. 3. (a) Dynamic Low-pass Filter. (b) Truncated Dynamic Low-pass Filter. Method (a) starts from zero which will lead to unreasonable inference results, on the contrary, method (b) ensures the validity of the inference information in the initial stage.



Truncated Dynamic Low-pass Filter：

$$\gamma_k(\mathbf{x}, \alpha) = \omega_k(\alpha) \cdot [\cos(2^k \pi \mathbf{x}), \sin(2^k \pi \mathbf{x})] \tag{7}$$

$$\omega_k(\alpha) = \begin{cases} 1 & k \leq (\alpha + \frac{\alpha_0}{L})L \\ 0 & k > (\alpha + \frac{\alpha_0}{L})L \end{cases} \tag{8}$$

1. Z. Zhu et al., "LATITUDE: Robotic Global Localization with Truncated Dynamic Low-pass Filter in City-scale NeRF," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 8326-8332, doi: 10.1109/ICRA48891.2023.10161570.

LATITUDE https://github.com/jike5/LATITUDE

清华大学车辆与运载学院
SCHOOL OF VEHICLE AND MOBILITY TSINGHUA UNIVERSITY

Scenes	Methods	Max	Mean	Min	Rmse	Std
UMAD	Regressor	7.03	1.69	0.34	2.07	1.20
	iNeRF	56.18	12.36	4.70	18.28	13.48
	Ours	**0.25**	**0.05**	**0.01**	**0.06**	**0.04**
Mill 19	Regressor	8.77	2.22	0.39	3.13	2.20
	iNeRF	33.81	17.52	19.42	20.43	10.51
	Ours	**0.15**	**0.11**	**0.06**	**0.11**	**0.02**

ABLATION STUDY WITH DIFFERENT INITIAL TRANSLATION ERROR.

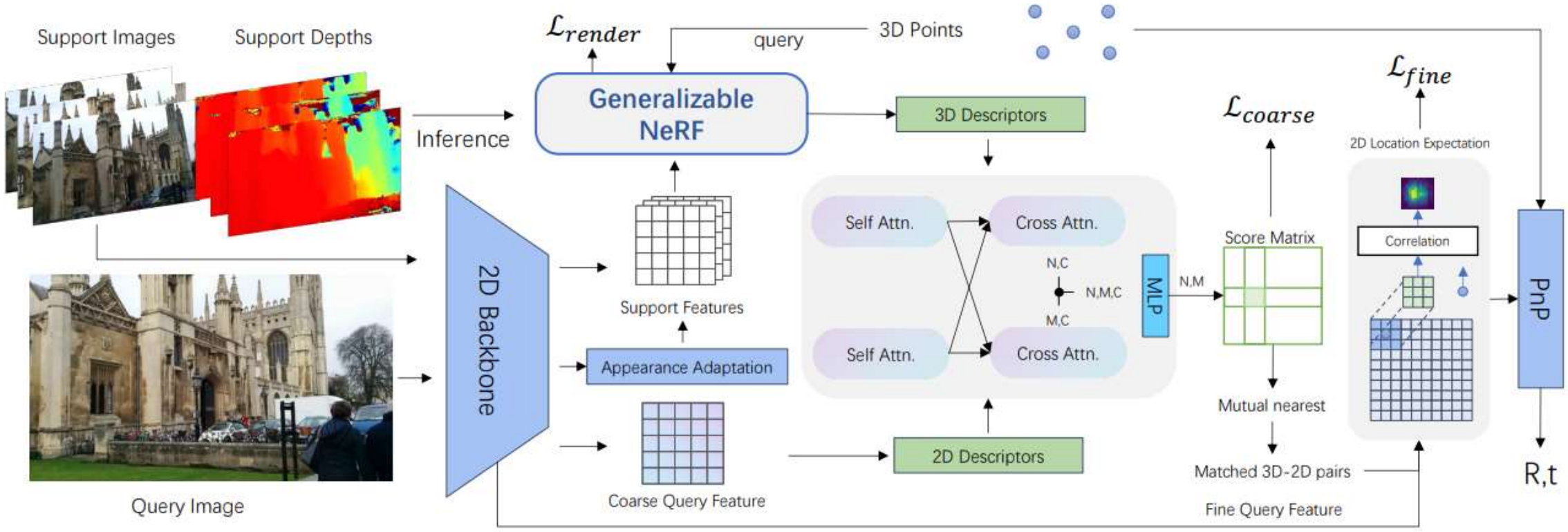
Initial Error(m)	Manifold Optimization	TDLF	Translation Error(m)	Rotation Error(°)
4	×	×	0.19	0.69
	✓	×	0.83	0.21
	✓	✓	**0.02**	**0.10**
8	×	×	22.78	8.81
	✓	×	1.97	0.17
	✓	✓	**0.02**	**0.09**
12	×	×	7.14	4.30
	✓	×	2.10	4.47
	✓	✓	**0.03**	**0.10**
16	×	×	7.41	5.86
	✓	×	4.19	2.10
	✓	✓	**3.84**	**0.97**

ABLATION STUDY WITH DIFFERENT INITIAL ROTATION ERROR.

Initial Error(°)	Manifold Optimization	TDLF	Translation Error(m)	Rotation Error(°)
4	×	×	8.72	13.5
	✓	×	0.86	1.03
	✓	✓	**0.02**	**0.09**
8	×	×	20.95	17.00
	✓	×	3.18	5.45
	✓	✓	**0.02**	**0.10**
12	×	×	7.81	28.92
	✓	×	6.12	11.03
	✓	✓	**0.70**	**3.39**
16	×	×	9.99	19.74
	✓	×	6.12	17.67
	✓	✓	**3.24**	**4.80**

1. Z. Zhu et al., "LATITUDE: Robotic Global Localization with Truncated Dynamic Low-pass Filter in City-scale NeRF," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 8326-8332, doi: 10.1109/ICRA48891.2023.10161570.

NeRF-Loc https://github.com/JenningsL/nerf-loc

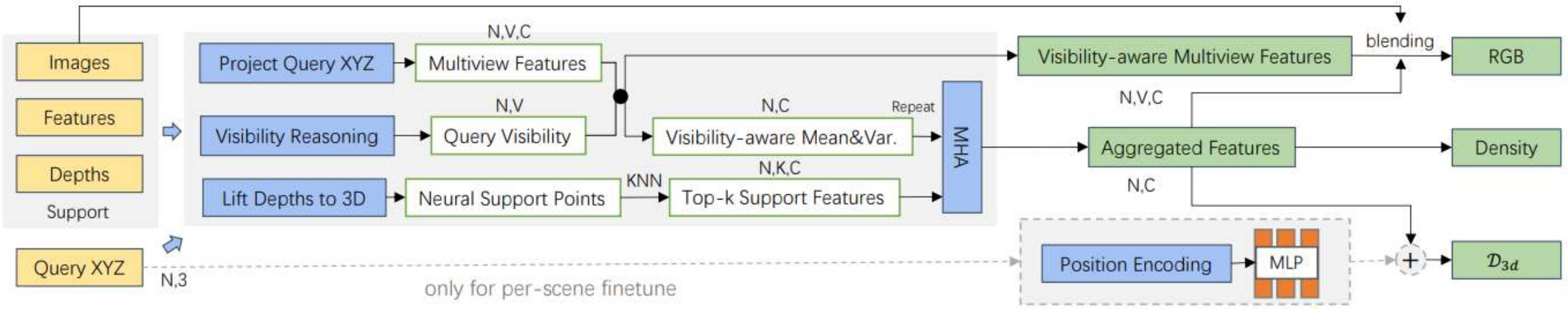


Support points: $P = \{P_s, F_s, \Lambda_s, D_s\}$

Loss: $\mathcal{L} = \mathcal{L}_{coarse} + \mathcal{L}_{fine} + \mathcal{L}_{render} + \mathcal{L}_{depth} \quad (2)$

1. J. Liu, Q. Nie, Y. Liu and C. Wang, "NeRF-Loc: Visual Localization with Conditional Neural Radiance Field," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 9385-9392, doi: 10.1109/ICRA48891.2023.10161420.

Conditional Neural 3D Model



$$\mathcal{M}(X) = \{f(x_i) = \sum_{k=0}^{K-1} \frac{w_{ik} * f'_{ik}}{K}; x_i \in X, w_i \in W, f'_i \in F'\} \quad (1)$$

$$W = W_a * W_d * \Lambda_s \in \mathbb{R}^{N \times K}$$

MHA Inverse distance Confidence

1. J. Liu, Q. Nie, Y. Liu and C. Wang, "NeRF-Loc: Visual Localization with Conditional Neural Radiance Field," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 9385-9392, doi: 10.1109/ICRA48891.2023.10161420.

Appearance Adaption Layer

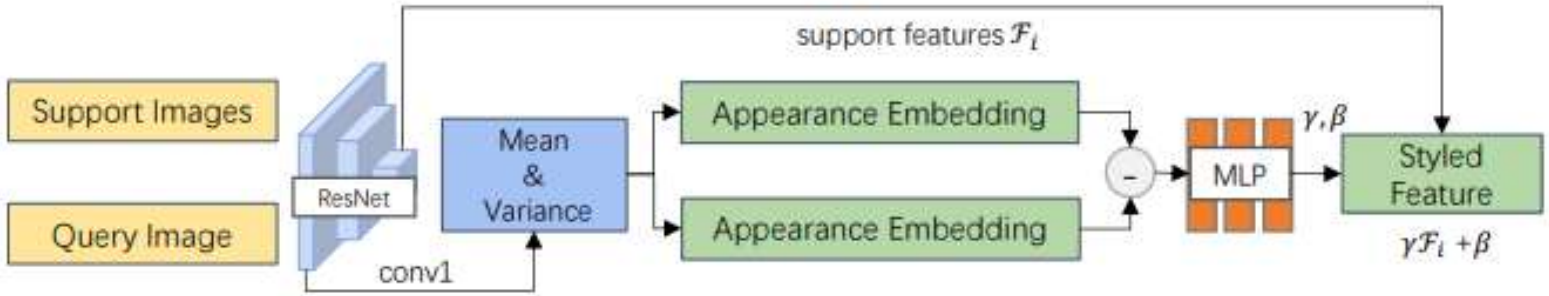


Fig. 3. Appearance adaptation.

Definition: Image(condition)

Image $\text{feature}_{s(s)}$ = $\text{conditioned features}_s$ + $\text{unconditioned features}_s$

Image $\text{features}_{q(q)}$ = $\text{conditioned features}_q$ + $\text{unconditioned features}_q$

Image $\text{feature}_{s(q)}$ = $\text{conditioned features}_q$ + $\text{unconditioned features}_s$

= ($\text{conditioned features}_q$ - $\text{conditioned features}_s$) + Image features_s

个人理解，不一定对

1. J. Liu, Q. Nie, Y. Liu and C. Wang, "NeRF-Loc: Visual Localization with Conditional Neural Radiance Field," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 9385-9392, doi: 10.1109/ICRA48891.2023.10161420.

NeRF-Loc https://github.com/JenningsL/nerf-loc

EVALUATION ON OUTDOOR AND INDOOR LOCALIZATION BENCHMARKS. WE REPORT THE MEDIAN TRANSLATION(CM) AND ROTATION(°) ERROR FOR EACH SCENE. **AVG.** STANDS FOR THE AVERAGE MEDIAN ERROR, AND **ACC.** IS THE ABBREVIATION FOR AVERAGE ACCURACY.

	Cambridge Landmarks - outdoor						7scenes - indoor							
Method	Church	Court	Hospital	College	Shop	Avg.↓	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Acc.
SANet[36]	16/0.57	328/1.95	32/0.53	32/0.54	10/0.47	83.6/0.8	3/0.9	3/1.1	2/1.5	3/1.0	5/1.3	4/1.4	16/4.6	68.2
DSAC[2]	55/1.6	280/1.5	33/0.6	30/0.5	9/0.4	81.4/0.9	2/0.7	3/1.0	2/1.3	3/1.0	5/1.3	5/1.5	190/49.4	60.2
InLoc[31]	18/0.6	120/0.6	48/1.0	46/0.8	11/0.5	48.6/0.7	3/1.1	3/1.1	2/1.2	3/1.1	5/1.6	4/1.3	9/2.5	66.3
DSM[32]	12/0.4	44/0.2	24/0.4	19/0.4	7/0.4	21.2/0.4	2/0.7	2/0.9	1/0.8	3/0.8	4/1.2	4/1.2	5/1.4	78.1
DSAC++[3]	13/0.4	40/0.2	20/0.3	18/0.3	6/0.3	19.4/0.3	**2/0.5**	2/0.9	1/0.8	3/0.7	4/1.1	4/1.1	9/2.6	74.4
DSAC*[4]	13/0.4	49/0.3	21/0.4	15/0.3	5/0.3	20.6/0.3	2/1.1	2/1.2	1/1.8	3/1.2	4/1.3	4/1.7	3/1.2	85.2
HACNet[13]	9/0.3	28/0.2	19/0.3	18/0.3	6/0.3	16.0/0.3	2/0.7	2/0.9	1/0.9	3/0.8	4/1.0	4/1.2	3/0.8	84.8
PixLoc[25]	10/0.3	30/0.1	**16/0.3**	14/0.2	5/0.2	15/0.2	2/0.8	**2/0.7**	**1/0.8**	3/0.8	4/1.2	**3/1.2**	5/1.3	75.7
Ours	**7/0.2**	**25/0.1**	18/0.4	**11/0.2**	**4/0.2**	**13/0.2**	2/1.1	2/1.1	1/1.9	**2/1.1**	**3/1.3**	3/1.5	**3/1.3**	**89.5**

1. J. Liu, Q. Nie, Y. Liu and C. Wang, "NeRF-Loc: Visual Localization with Conditional Neural Radiance Field," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 9385-9392, doi: 10.1109/ICRA48891.2023.10161420.