FDC-NeRF: Learning Pose-Free Neural Radiance Fields with Flow-Depth Consistency

Huachen Gao, Shihe Shen, Zhe Zhang, Kaiqiang Xiong, Rui Peng, Zhirui Gao,

Qi Wang, Yugui Xie, Ronggang Wang*

(*corresponding author)

PEKING UNIVERSITY 1898

Introduction

Neural Radiance Fields (NeRF) is a method for reconstructing the 3D model of objects or scenes from multiple input images. NeRF enables high-fidelity Novel View Synthesis.

Downstream: Autonomous Driving, 3D Generations, Augmented Reality & Virtual Reality, Metaverse, etc.

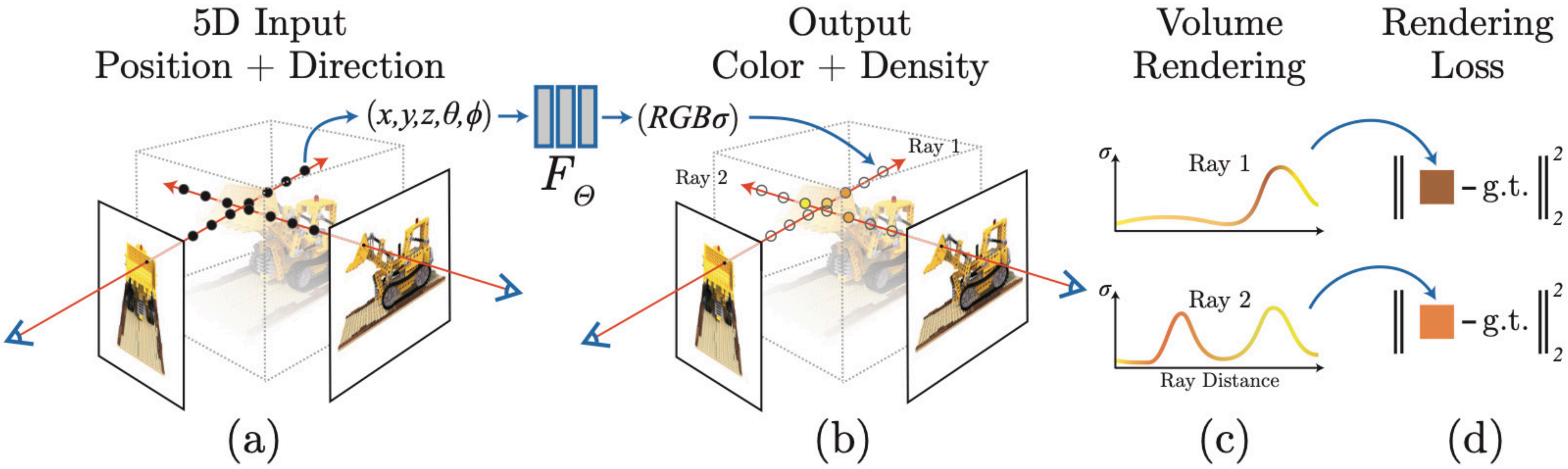


Fig: NeRF architecture.

Introduction

Crucial Requirement for NeRF

- Reliable annotated camera-parameters.
- Camera estimation (Stucture-from-Motion) pre-processing, e.g. COLMAP
- SfM often fails in low-textured areas, few overlapping views, occlusions…

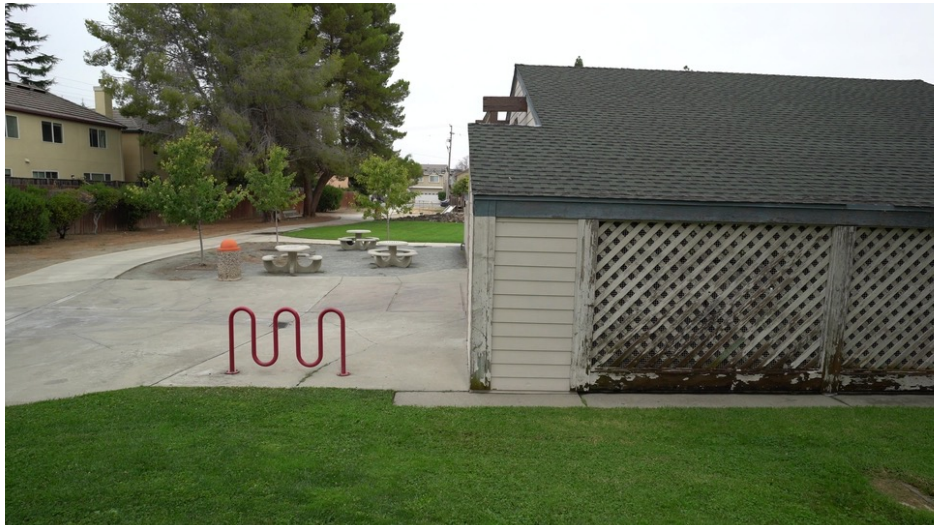
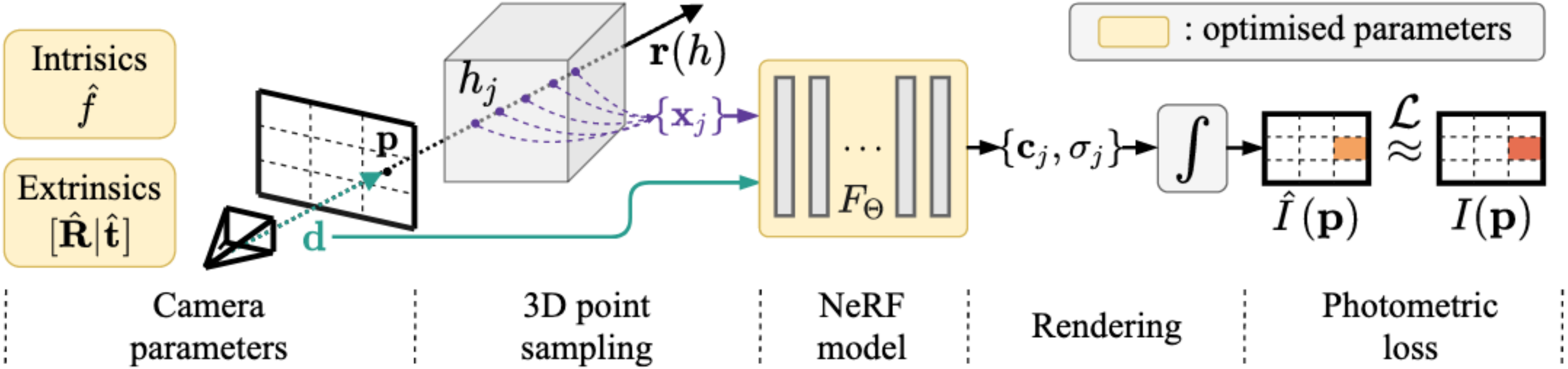


Fig: Scenes with low-texture, few overlapping and occlusions.

Introduction

Solution: Joint Estimation of Camera Poses and NeRF

- **NeRFmm: Neural Radiance Fields Without Known Camera Parameters**
- **BARF : Bundle-Adjusting Neural Radiance Fields**



BARF additionally proposed a coarse-to-fine positional encoding for joint learning.

[1] Wang, Z., Wu, S., Xie, W., Chen, M., & Prisacariu, V. (2021). NeRF$--$: Neural Radiance Fields Without Known Camera Parameters. *Cornell University - arXiv,Cornell University - arXiv.*
[2] Lin, Chen-Hsuan, et al. "Barf: Bundle-adjusting neural radiance fields." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Introduction

Problems

- Previous methods relies heavily on **pose initialization**.
- Can not deal with unbounded scenes with **large camera movements**.

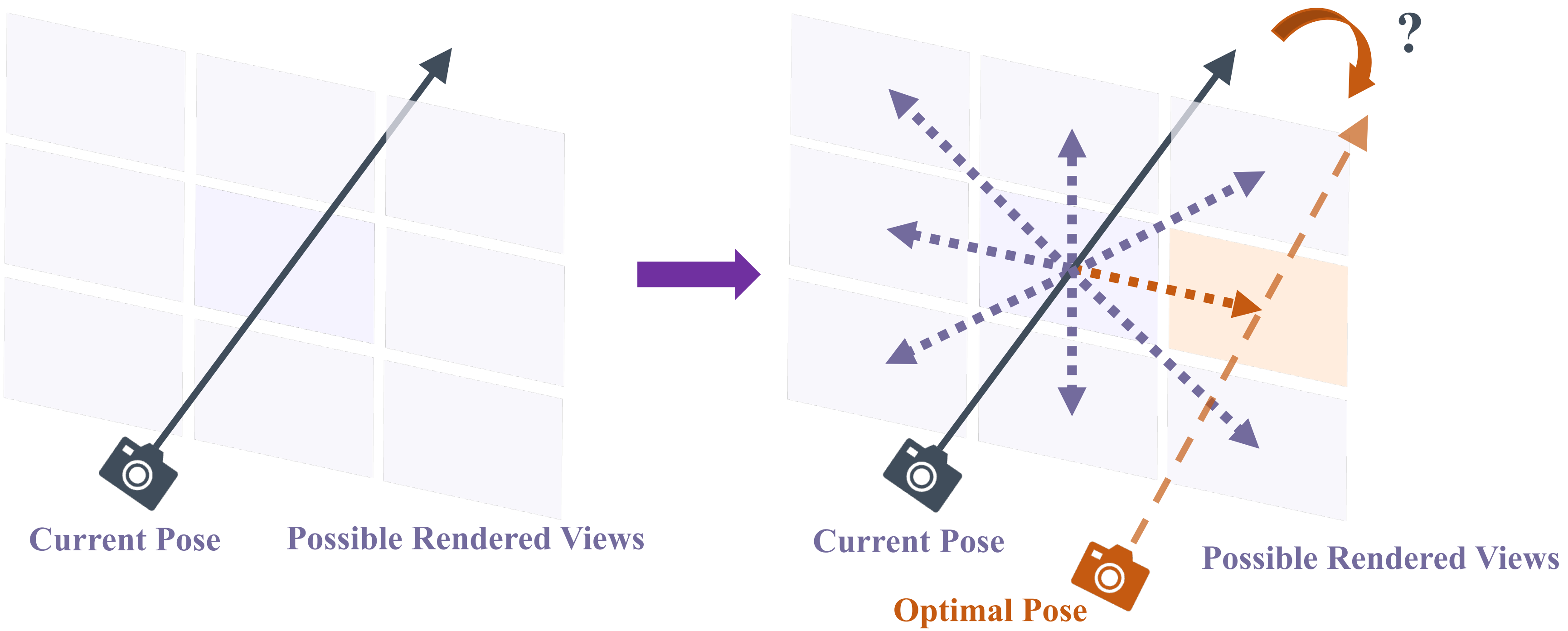
Concurrent Solution:

- **NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior**
- Deal with unbounded scenes of large camera movement **without** any pose prior.
- Integrated with mono-depth estimation as geometry regularization.
- Relative pose constraint in consecutive frames and 3D point cloud supervision.

[1] Bian, Wenjing, et al. "Nope-nerf: Optimising neural radiance field with no pose prior." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Motivation

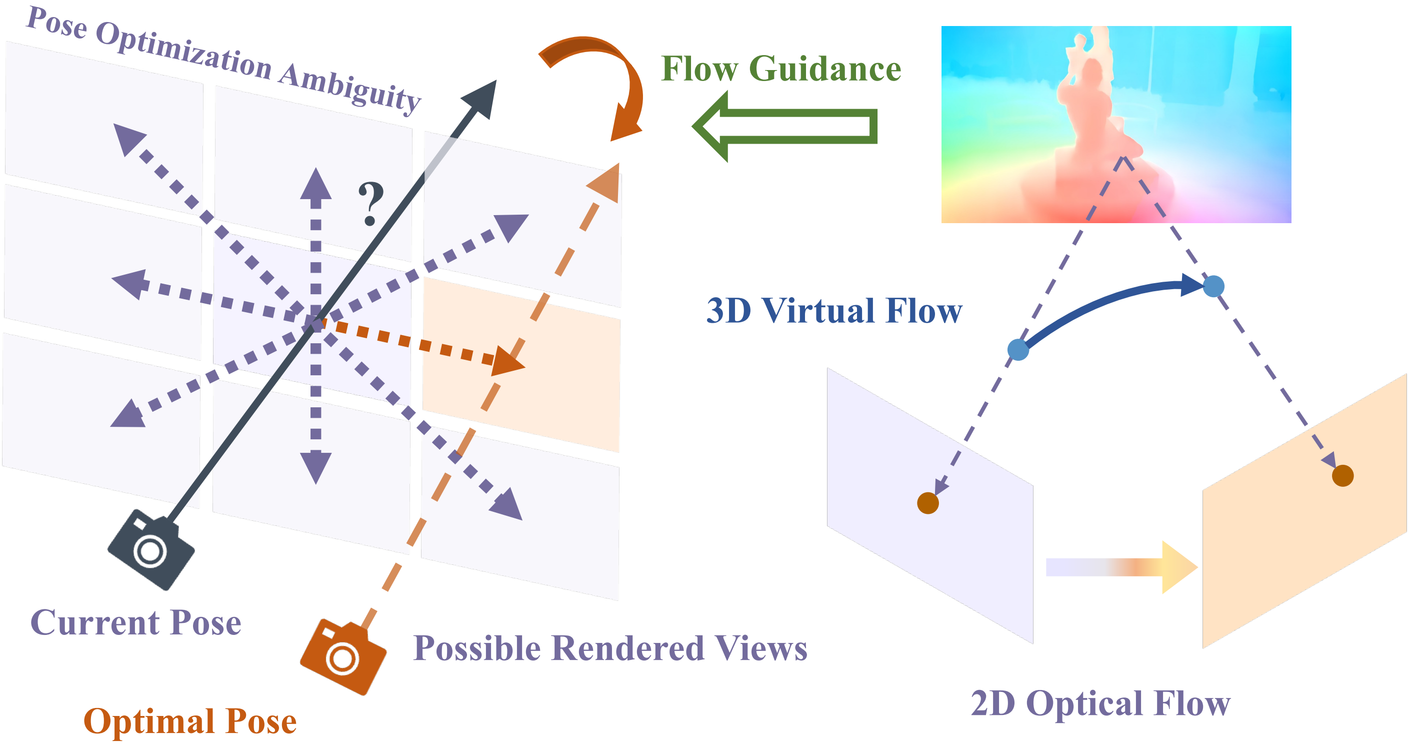
The key challenge lies in the pose-free NeRF is the <u>**pose-geometry ambiguity**</u>



Step 1 … **Step N+1**

NeRF's geometric ambiguity leads to uncertain gradients for pose optimization

Our Solution



“ Leverage the **direction information** embedded in the 2D optical flow to provide **direct guidance** for pose optimization. ”

Specifically, we enforce **RGB-based optical flow** to be consistent with **depth-based virtual flow** to excavate geometry clues across views.

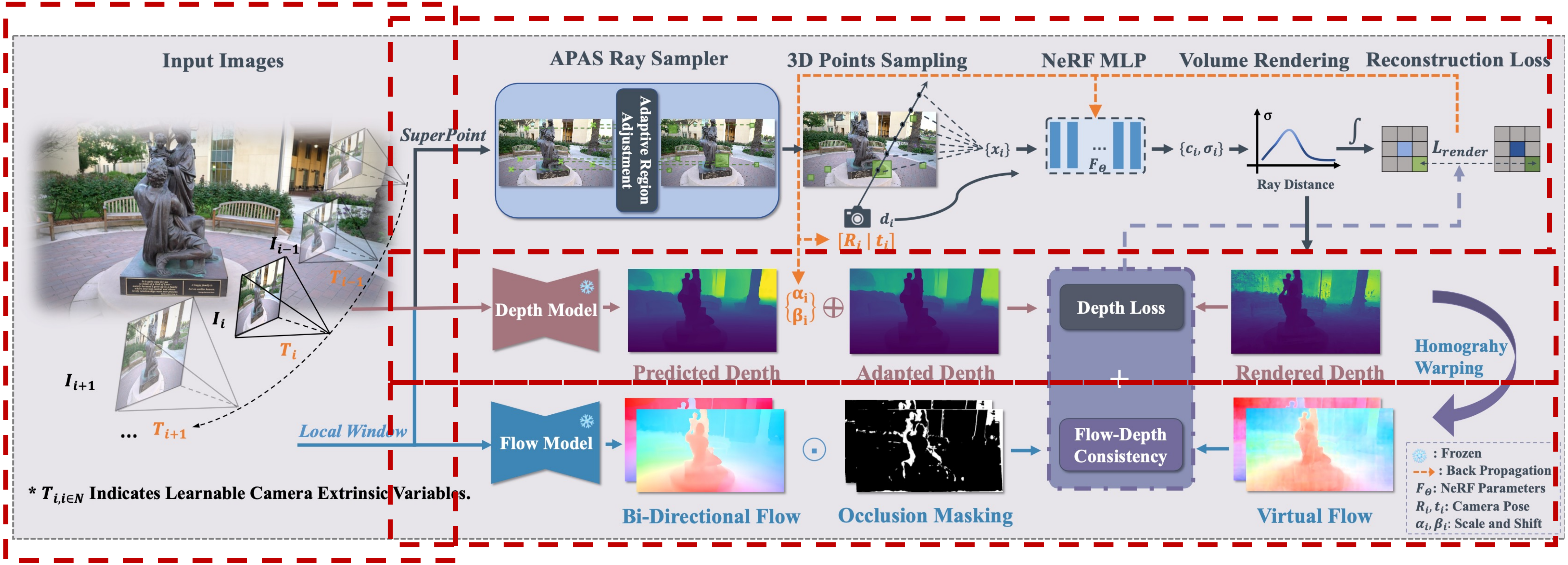
Contributions

FDC-NeRF: a novel pose-free NeRF for unbounded scenes with large camera movement

1. We propose the **Flow-Depth Consistency Guidance**, which leverages the direction information in 2D optical flow and virtual flow to provide direct guidance for joint pose-NeRF optimization.
2. We introduce the **Adaptive Pose-Aware Sampling (APAS)** strategy to sample pose-aware feature points for more effective pose supervision, and adaptively adjust the sampling regions to increase the diversity of rays.
3. Experimental results on Tanks and Temples dataset with large camera movements show that our method achieves state-of-the-art performance on **both novel view synthesis quality and pose estimation accuracy**.

Method

Overall Pipeline



* $T_{i,i\in N}$ Indicates Learnable Camera Extrinsic Variables.

Method

Flow-Depth Consistency Guidance

[Monodepth Regularization]

We follow NoPe-NeRF to regularize NeRF rendered depth with adapted monocular depth estimation to enhance the representation of scene geometry, making it less prone to get stuck in local minima.

We generate monocular depth sequence by DPT $\{D_i\}_{i=1}^{N}$

Then, we build learnable scale and shift parameters to linearly adapt the mono-depth maps $\Phi = \{\alpha_i, \beta_i\}_{i=1}^{N}$

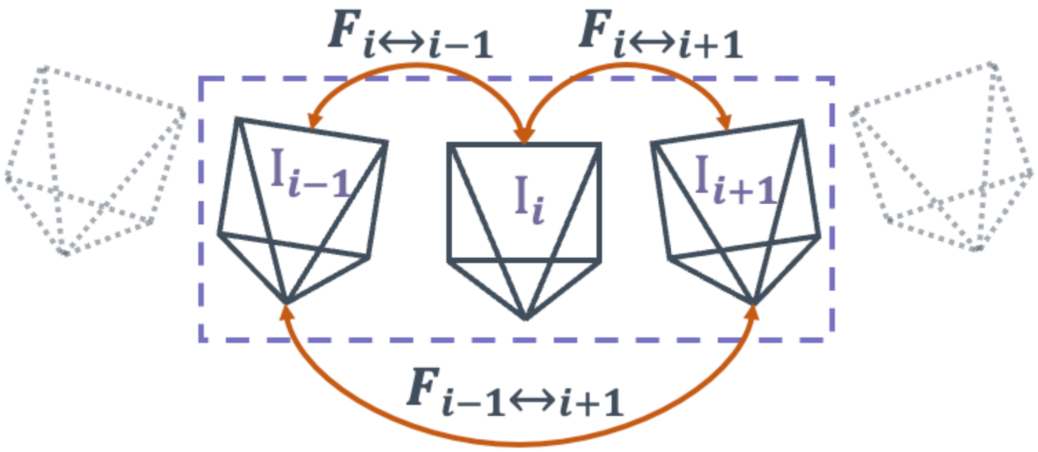
The regularization between rendered depth and mono-depth can be formulated as

$$\mathcal{L}_{Depth} = \Sigma_{i=1}^{N} ||(\alpha_i D_i + \beta_i) - \hat{D}_i||$$

Method

Flow-Depth Consistency Guidance

[local window strategy]



Suppose we are conducting Flow-Depth Consistency Guidance on frame i, the local window consists of the frame i-1 and frame i+1.

The regularization is performed between consecutive frames (i with i+1 and i-1) and cross-view frames (i-1 with i+1) to **reduce errors caused by the misaligned unidirectional flow.**

Method

Flow-Depth Consistency Guidance

[How Flow-Depth Consistency works]

Step1: Generating 2D Optical Flow

We utilize pretrained GMFlow to estimate bidirectional optical flow between frames in RGB level. $\mathbf{F_{i \leftrightarrow i+1}}$

Run forward-backward consistency check to obtain the occlusion masks. $M_{i \rightarrow i+1} = \{|\mathbf{F_{i \rightarrow i+1}} + \mathbf{F_{i+1 \rightarrow i}}| > 0.5\}$

Step2: Generating 3D Virtual Flow

Calculate homography warping w.r.t. NeRF rendered depth and estimated pose

$$\Pi_{i \rightarrow i+1} = K_{i+1} T_{i+1} T_i^{-1} K_i^{-1} \hat{D}_i$$

Method

☐ Flow-Depth Consistency Guidance

[How Flow-Depth Consistency works]

Step3: Enforce 2D Optical Flow consistent with 3D Virtual Flow

Denote (u_k, v_k) the 2D coordinate value of a pixel p_k, the forward virtual flow can be formulated as

$$\hat{\mathbf{F}}_{i\to i+1} = \Pi_{i\to i+1}((u_k, v_k)) - (u_k, v_k), (p_k \in I_i)$$

We construct forward consistency between RGB- based optical flow and virtual flow on the non-occluded valid region

$$\mathcal{L}_{Flow}^{i\to i+1} = \frac{||(\hat{\mathbf{F}}_{i\to i+1} - \mathbf{F}_{i\to i+1}) \odot M_{i\to i+1}||_2}{||M_{i\to i+1}||_1}$$

Method

Flow-Depth Consistency Guidance

[How Flow-Depth Consistency works]

Step4: Conduct Guidance in Local Window

$$\mathcal{L}_{Flow} = w_1\mathcal{L}_{Flow}^{i\leftrightarrow i+1} + w_2\mathcal{L}_{Flow}^{i\leftrightarrow i-1} + w_3\mathcal{L}_{Flow}^{i-1\leftrightarrow i+1}$$

we set $w_1 = 0.4, w_2 = 0.4, and\ w_3 = 0.2$. The camera poses and rendered depth are directly used for homography warping

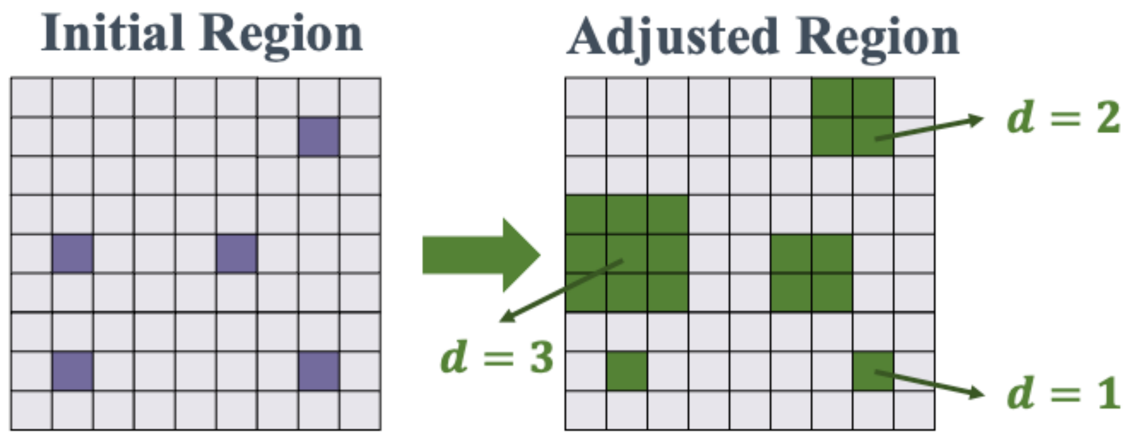
Method

☐ Adaptive Pose-Aware Sampling Strategy

[Why APAS]

Recent NeRFs apply the **random ray sampling** strategy while increasing the probability of **sampling in low-texture regions**. These rays exhibit **indistinguishable photometric error**, which potentially **aggravates pose-geometry ambiguity**.

[How APAS works]



$\mathcal{N}_i$ Expand Based on Photometric Error to $\hat{\mathcal{N}_i}$

The SuperPoint is first adopted to obtain a set of feature points $\{p_1, p_2, \ldots p_M\}$ with pose-aware features for the initial samplings.

We then refer the idea of curriculum learning to adapt sampling region according to photometric error on each ray for the subsequent ray selection.

$$\hat{\mathcal{N}_i} = \mathcal{N}_i \bigcup [d_i + d_{max} \times \sigma(log\frac{\mathcal{L}^i_{Render}}{\mathcal{L}_{Render}})]^2$$

Experiments

Tanks and Temples: outdoor, long sequence, unbounded real scenes

NeRFmm | **BARF** | **NoPe-NeRF** | **Ours** | **Ground Truth**

Horse
Ballroom
Ignatius
Barn

Experiments

Qualitative

Horse

Ignatius

NoPe-NeRF CVPR 2023 **Ours**

Experiments

Quantitative on NVS

Table 1: Quantitative results of novel view synthesis on Tanks and Temples [13] dataset. Bold numbers represent the best. Each method is trained with public code and original parameters, and evaluated under the same settings.

Scenes	Ours			NoPe-NeRF [10]			BARF [5]			NeRFmm [4]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Church	**25.53**	**0.74**	**0.37**	25.17	0.73	0.39	23.17	0.62	0.52	21.64	0.58	0.54
Barn	**26.86**	**0.71**	**0.40**	26.33	0.69	0.44	25.28	0.64	0.48	23.21	0.61	0.53
Museum	**26.90**	**0.77**	**0.33**	26.41	**0.77**	0.36	23.58	0.61	0.55	22.37	0.61	0.53
Family	**26.51**	**0.76**	**0.39**	26.01	0.74	0.41	23.04	0.61	0.56	23.04	0.58	0.56
Horse	**27.73**	**0.84**	**0.25**	26.58	0.81	0.29	24.09	0.72	0.41	23.12	0.70	0.43
Ballroom	**25.65**	**0.74**	**0.35**	25.16	0.69	0.41	20.66	0.50	0.60	20.03	0.48	0.57
Francis	**28.92**	**0.79**	**0.39**	28.54	0.78	0.41	25.85	0.69	0.57	25.40	0.69	0.52
Ignatius	**24.26**	**0.61**	0.48	23.81	**0.61**	**0.47**	21.78	0.47	0.60	21.16	0.45	0.60
mean	**26.55**	**0.75**	**0.37**	26.01	0.73	0.40	23.43	0.60	0.54	22.49	0.58	0.53

Experiments

Quantitative on Pose Estimation

Table 2: Quantitative results of pose estimation on Tanks and Temples [13] dataset. Best results are **Bolded**. We take the COLMAP estimation as the ground truth poses.

Scenes	Ours			NoPe-NeRF [10]			BARF [5]			NeRFmm [4]		
	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE↓	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE↓	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE↓	$RPE_t \downarrow$	$RPE_r \downarrow$	ATE↓
Church	**0.030**	**0.006**	**0.004**	0.045	0.008	0.008	0.458	0.063	0.059	0.626	0.127	0.065
Barn	**0.026**	**0.019**	**0.004**	0.033	0.029	**0.004**	1.402	0.326	0.075	1.629	0.494	0.159
Museum	**0.207**	0.245	**0.021**	**0.207**	**0.240**	0.023	2.589	1.128	0.257	4.134	1.051	0.346
Family	**0.031**	**0.009**	**0.002**	0.047	0.011	**0.002**	0.577	0.595	0.116	2.743	0.537	0.120
Horse	0.188	**0.018**	0.004	**0.179**	0.039	**0.003**	0.239	0.399	0.016	1.349	0.434	0.018
Ballroom	**0.034**	**0.019**	**0.001**	0.058	0.025	0.003	0.343	0.228	0.019	0.449	0.177	0.031
Francis	**0.063**	**0.031**	**0.005**	0.079	0.042	0.011	0.924	0.749	0.095	1.647	0.618	0.207
Ignatius	**0.029**	0.007	**0.002**	0.037	**0.006**	0.004	1.187	0.288	0.057	1.302	0.379	0.041
mean	**0.076**	**0.044**	**0.005**	0.085	0.050	0.007	0.965	0.472	0.087	1.735	0.477	0.123

Experiments

Ablation Study

Table 3: Ablation results on Tanks and Temples [13].

Methods	NVS			Pose Est.		
	PSNR↑	SSIM↑	LPIPS↓	RPE_t ↓	RPE_r ↓	ATE↓
Ours	**26.55**	**0.75**	**0.37**	**0.076**	**0.044**	**0.005**
Ours w/o $\mathcal{L}_{Flow}$	25.84	0.72	0.40	0.101	0.057	0.011
Ours w/o APAS	26.20	0.73	**0.37**	0.086	0.051	0.007
Ours w/o $\mathcal{L}_{Flow}$ and APAS	25.35	0.70	0.41	0.124	0.062	0.015

1) Flow-Depth Consistency Guidance: When the flow-depth consistency is not considered, due to the absence of direct guidance, it is more difficult for pose-free NeRF to optimize effectively and leading to pose-geometry ambiguity.

2) Adaptive Pose Aware Sampling: When disabling the APAS strategy, the rays sampled by random sampling may provide less effective supervision compared to APAS, resulting in lower performance in pose accuracy and synthesis quality.

Experiments

Limitations and Future Works

1) Long training time. Requires 25-28 hours to train for one scene.
2) Sequence of input. Requires video sequence, the pretrained optical flow require consecutive frames.

Future Works

1) Apply 3D Gaussian Splatting into joint estimation for faster training and rendering speed.
2) Reduce the input views to sparse views input.

Thanks for Listening!

Huachen Gao, Shihe Shen, Zhe Zhang, Kaiqiang Xiong, Rui Peng, Zhirui Gao,

Qi Wang, Yugui Xie, Ronggang Wang*

(*corresponding author)