# FDC-NERF: LEARNING POSE-FREE NEURAL RADIANCE FIELDS WITH FLOW-DEPTH CONSISTENCY

*Huachen Gao*[1], *Shihe Shen*[1], *Zhe Zhang*[1], *Kaiqiang Xiong*[1],
*Rui Peng*[1], *Zhirui Gao*[2], *Qi Wang*[3], *Yugui Xie*[3], *Ronggang Wang*[1*]

[1]School of Electronic and Computer Engineering, Peking University, China
[2]National University of Defense Technology, China [3]MIGU Video Co., Ltd., China

## ABSTRACT

Learning neural radiance fields (NeRF) without camera poses has been widely studied. However, recent methods lack explicit and effective supervision for pose estimation, resulting in ambiguous optimization of camera pose and NeRF geometry during joint training, particularly in scenarios involving large camera movements. In this paper, we propose FDC-NeRF that leverages the direction information contained in the RGB-based optical flow and depth-based virtual flow as a direct guidance for camera pose optimization to reduce pose-geometry ambiguity. Additionally, we introduce Adaptive Pose-Aware Sampling (APAS) to replace the previous random ray sampling strategy, which reduces the difficulty of pose learning in early stages and preserves the diversity of rays in later stages. Experiments on the challenging Tanks and Temples dataset demonstrate that our method achieves state-of-the-art results in both novel view synthesis quality and pose estimation accuracy.

***Index Terms***— Neural Radiance Fields, Pose Estimation, Novel View Synthesis, 3D Reconstruction

## 1. INTRODUCTION

Neural Radiance Fields (NeRF) [1] has demonstrated powerful capability in 3D scene representation and high-fidelity Novel View Synthesis (NVS), which is widely applied in VR/AR, 3D content generation, etc. A crucial prerequisite for NeRF reconstruction is the reliable annotated camera parameters. However, accurate poses are not always feasible in real-world scenarios, and the pre-processing of camera parameters heavily relies on offline methods like Structure-from-Motion (SfM) [2].

*Ronggang Wang is the corresponding author (rgwang@pkusz.edu.cn).

This work is financially supported by National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Science and Technology Program-Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents project (Grant No. RCJC20200714114435057), Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project (Grant No. SGDX20211123144400001), this work is also financially supported for Outstanding Talents Training Fund in Shenzhen.

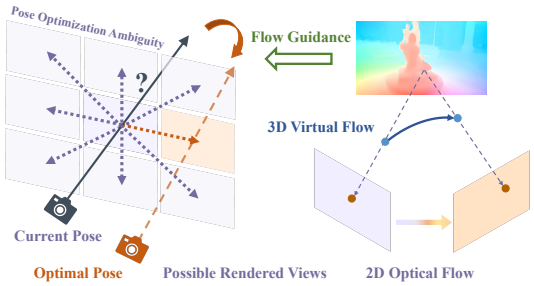



**Fig. 1**: **Illustration of pose-geometry ambiguity and flow-guided pose optimization.**

To reduce the reliance on camera parameters, efforts have been made to estimate camera poses and radiance fields simultaneously. In particular, iNeRF [3] utilizes reconstructed NeRF to estimate camera poses for novel viewpoints. NeRFmm [4] proposes an end-to-end pipeline to jointly estimate camera extrinsics, intrinsic, and NeRF. BARF [5] analyzes the connection between planer image alignment and NeRF, and proposes a coarse-to-fine positional encoding. SC-NeRF [6] further estimates camera distortion with geometric regularization. GARF [7] and SiNeRF [8] facilitate easier joint learning by adopting different activation functions in NeRF MLPs. These methods are limited to pose initialization close to the ground truth. GNeRF [9] adopts GAN to reduce this limitation but still requires a known pose sampling distribution. Recently, NoPe-NeRF [10] integrates mono-depth maps into joint estimation to tackle the challenging large camera movements without pose priors, which is more relevant to our work. However, these methods still fail to recover adequate scene geometry for accurate pose estimation.

The key challenge lies in the pose-free NeRF is the pose-geometry ambiguity. NeRF tends to overfit 3D scenes for photorealistic RGB restoration but lacks explicit geometry learning [11], while accurate camera pose estimation relies on abundant geometry guidance. During joint optimization, NeRF's geometric ambiguity leads to uncertain gradients for pose optimization and inaccurate pose estimation results in poor NeRF reconstruction. This ambiguity grows when handling large camera movements. To mitigate the dependency of pose learning on NeRF geometry, as shown in Fig.1, we

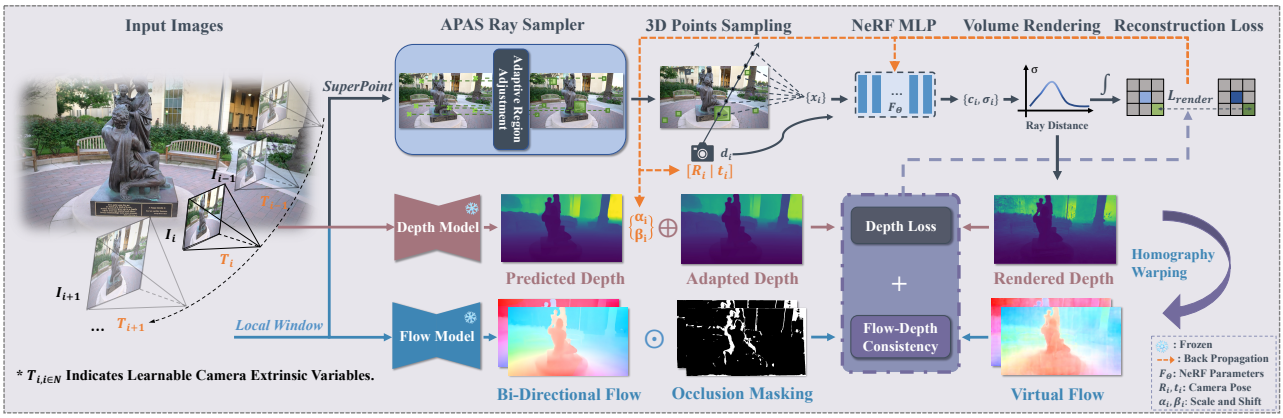



**Fig. 2**: **Network architecture.** In our pipeline, NeRF MLP and the parameters in orange are jointly optimized, where the novel APAS is proposed for ray sampling. We introduce mono-depth loss and propose flow-depth consistency for joint optimization.

propose the Flow-Depth Consistency Guidance which leverages 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. Additionally, the random sampling strategy applied in previous methods may introduce less anisotropic supervision in low-textured regions, which further increases the pose-geometry ambiguity. We draw inspiration from curriculum learning [12] and introduce the Adaptive Pose-Aware Sampling (APAS) strategy. The APAS starts by extracting pose-aware feature points for effective supervision in pose estimation, then adaptively adjusts the sampling regions to broaden the diversity of rays for better scene representation.

In summary, the primary contributions are as follows:

- 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.
- We introduce the 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.
- Experimental results on Tanks and Temples dataset [13] with large camera movements show that our method achieves state-of-the-art performance on both novel view synthesis quality and pose estimation accuracy.

## 2. METHODOLOGY

We provide an overview pipeline in Fig. 2. Given a sequence of images $\{I_i\}_{i=1}^N$ with large camera movements and camera intrinsics $\{K_i\}_{i=1}^N$, our goal is to reconstruct the radiance field of the 3D scene **without** camera extrinsics $\{T_i\}_{i=1}^N$[4]. We take NeRFmm [4] integrated with monocular depth supervision as our baseline, and propose the flow-depth consistency guidance (Sec.2.2) with the APAS strategy (Sec.2.3) to tackle the challenge of handling large camera movements without good pose initialization.

### 2.1. Formulation and Preliminaries

NeRF represents scene as a view-dependent mapping function $F_\Theta : (x, d) \rightarrow (c, \sigma)$ parameterized by an MLP $\Theta$, where $x \in \mathbb{R}^3$ is the 3D point location, $d \in \mathbb{R}^3$ is the corresponding viewing direction, $c \in \mathbb{R}^3$ and $\sigma \in \mathbb{R}$ are the radiance color and density values respectively. A synthesized image $\hat{I}$ can be rendered by compositing radiance color and density along camera rays $r(t) = o + td$ between near and far plane $t_n$ and $t_f$. The volume rendering [14, 15] is formulated as:

$$\hat{I}_i(\mathbf{r}) = \int_{t_n}^{t_f} W(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt, \tag{1}$$

where $W(t) = exp(-\int_{t_n}^{t} \sigma(r(s))\, ds)$ is accumulated transmittance along the ray, and $W(t)$ is utilized to render depth as $\hat{D}_i = \int_{t_n}^{t_f} W(t)\sigma(r(t))\, dt$. The NeRF parameters $\Theta$ and the camera pose $T_i$ that are used to cast the ray can be jointly estimated by minimizing the photometric loss:

$$\mathcal{L}_{Render} = \Sigma_{i=1}^{N} ||I_i - \hat{I}_i||_2^2. \tag{2}$$

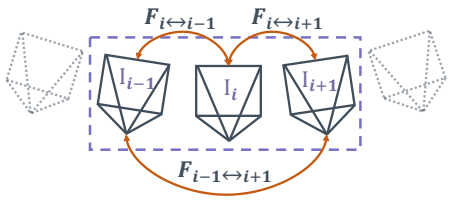
Besides, we follow [10, 16] to regularize NeRF rendered depth with adapted monocular depth estimation [17, 18] to enhance the representation of scene geometry, making it less prone to get stuck in local minima. Specifically, we generate monocular depth sequence $\{D_i\}_{i=1}^N$ by DPT [17] , and build learnable *scale* and *shift* parameters $\Phi = \{\alpha_i, \beta_i\}_{i=1}^N$ to linearly adapt the mono-depth maps. The regularization between rendered depth and mono-depth can be formulated as:

$$\mathcal{L}_{Depth} = \Sigma_{i=1}^{N} ||(\alpha_i D_i + \beta_i) - \hat{D}_i||, \tag{3}$$

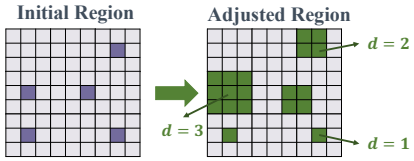
where $\Phi$ is jointly estimated for multi-view consistency.

### 2.2. Flow-Depth Consistency Guidance

As discussed in Sec.1, we address the pose-geometry ambiguous problem by exploiting the consistency between optical flow and 3D virtual flow to leverage direction information for pose-free NeRF optimization. First, we propose a local window strategy to impose flow-depth consistency between different viewpoints, providing more effective guidance for the optimization of camera parameters. Denote an image $I_i, i \in [2, N-1]$, a local window consists of $I_i$ and its




**(a) Local Window for Flow-Depth Consistency**




**(b) Adaptive Sampling Region Adjustment**

**Fig. 3**: **Illustrations of our proposed methods.** (a) Flow-depth consistency is conducted in the Local Window. (b) Single-step of adaptive region adjustment, where $d$ is calculated based on the photometric loss along each ray.

adjacent frames $\{I_{i-1}, I_{i+1}\}$. As shown in Fig.3(a), our regularization is similarly performed between consecutive frames $I_i \leftrightarrow \{I_{i-1}, I_{i+1}\}$ and cross-view frames $I_{i-1} \leftrightarrow I_{i+1}$ to reduce errors caused by the misaligned unidirectional flow. The following discussion focuses on $I_i$ and $I_{i+1}$ for simplicity.

We utilize GMFlow [19] to estimate bidirectional optical flow $\mathbf{F_{i\leftrightarrow i+1}}$ between frames in RGB level. The occluded parts are masked out by running forward-backward consistency check [20] as $M_{i\rightarrow i+1} = \{|\mathbf{F_{i\rightarrow i+1}} + \mathbf{F_{i+1\rightarrow i}}| > 0.5\}$ to prevent the learning of incorrect deformation in the occluded pixels. The 3D virtual flow is the coordinate offsets of corresponding points between different views, which is generated by projecting one viewpoint to another using depth and camera parameters. The virtual flow is generated as follows and enforced to be consistent with RGB-based optical flow. Denote a pixel $p_k$ on $I_i$, the corresponding pixel on $I_{i+1}$ can be obtained via differentiable homography warping $\Pi$ [21] w.r.t. rendered depth $\hat{D}_i$ and estimated pose $T_i$ and $T_{i+1}$:

$$\Pi_{i\rightarrow i+1} = K_{i+1} T_{i+1} T_i^{-1} K_i^{-1} \hat{D}_i, \tag{4}$$

where $T_{i+1}T_i^{-1}$ represents relative camera poses, bringing a pixel in the $i^{th}$ camera's space to the $i+1^{th}$'s. Denote $(u_k, v_k)$ the 2D coordinate values of pixel $p_k$, the forward virtual flow $\hat{\mathbf{F}}_{i\rightarrow i+1}$ from $I_i$ to $I_{i+1}$ can be formulated as:

$$\hat{\mathbf{F}}_{i\rightarrow i+1} = \Pi_{i\rightarrow i+1}((u_k, v_k)) - (u_k, v_k), (p_k \in I_i). \tag{5}$$

Finally, we construct forward consistency between RGB-based optical flow and virtual flow on the non-occluded valid region from $I_i$ to $I_{i+1}$, which can be calculated as:

$$\mathcal{L}_{Flow}^{i\rightarrow i+1} = \frac{||(\hat{\mathbf{F}}_{i\rightarrow i+1} - \mathbf{F}_{i\rightarrow i+1}) \odot M_{i\rightarrow i+1}||_2}{||M_{i\rightarrow i+1}||_1}, \tag{6}$$

where $\odot$ denotes the point-wise product, we apply a similar process to regularize the backward flow $\mathcal{L}_{Flow}^{i+1\rightarrow i}$, then the flow-depth consistency $\mathcal{L}_{Flow}^{i\leftrightarrow i+1}$ between two frames is the average of the forward's and backward's. The final calculation of flow-depth consistency in a local window is:

$$\mathcal{L}_{Flow} = w_1 \mathcal{L}_{Flow}^{i\leftrightarrow i+1} + w_2 \mathcal{L}_{Flow}^{i\leftrightarrow i-1} + w_3 \mathcal{L}_{Flow}^{i-1\leftrightarrow i+1}, \tag{7}$$

where we set $w1 = 0.4, w_2 = 0.4$ and $w_3 = 0.2$. The camera poses and rendered depth are directly used for homography warping in Equ.5, the 2D direction between frames embedded in optical flow and virtual flow gives a much clearer 3D gradient direction for joint optimization.

### 2.3. Adaptive Pose-Aware Sampling Strategy

Recent NeRFs apply the random ray sampling strategy while increasing the probability of sampling in low-texture regions. The rays sampled in these regions exhibit indistinguishable photometric error, which potentially aggravates pose-geometry ambiguity. To tackle this issue, inspired by [3], we propose a novel sampling strategy that adaptively adjusts the sample areas from pose-aware positions to the entire image. The SuperPoint [22] is first adopted to obtain a set of feature points $\{p_1, p_2, ...p_M\}$ with pose-aware features for the initial samplings. These points provide more effective supervision for pose estimation when the scene geometry and camera poses are ambiguous.

Denote $p_i$ in the feature point set, and the current sampling region around $p_i$ is $\mathcal{N}_i$, we then refer the idea of curriculum learning [12] to adapt the region to $\hat{\mathcal{N}}_i$ for the subsequent random ray selection. Specifically, we expand $\mathcal{N}_i$ based on the contribution of the photometric loss $L_{Render}^i$ on $p_i$ to the overall loss $L_{Render}$ generated by all rays, as a low photometric error indicates sufficient learning of 3D scene along this ray, the sampling region will be expanded to a larger extent. We expand $\mathcal{N}_i$ and conduct random sampling in $\hat{\mathcal{N}}_i$ to increase the complexity of the follow-up training, enabling NeRF to capture additional scene details. Assuming that $\mathcal{N}_i$ is $d_i \times d_i$, as shown in Fig.3(b), the adjustment of region $\mathcal{N}_i$ is as follow:

$$\hat{\mathcal{N}}_i = \mathcal{N}_i \bigcup [d_i + d_{max} \times \sigma(log \frac{\mathcal{L}_{Render}^i}{\mathcal{L}_{Render}})]^2. \tag{8}$$

Here, $\sigma$ represents the sigmoid function and $d_{max} = 5$ indicates the maximum range for adjustment. The region adjustment is conducted every 200 iterations until different regions are merged to encompass the entire image. Finally, the APAS returns to random sampling to increase ray diversity.

### 2.4. Loss Function

We assemble all constraints mentioned in previous sections, and formulate the total loss as:

$$Loss = \mathcal{L}_{Render} + \lambda_1 \mathcal{L}_{Depth} + \lambda_2 \mathcal{L}_{Flow}, \tag{9}$$

where we set $\lambda_1 = 0.04, \lambda_2 = 0.01$ in our experiments. We jointly optimize camera poses $T$, NeRF parameters $\Theta$, *scale* and *shift* parameters $\Phi$ by minimizing the above loss as:

$$T^*, \Theta^*, \Phi^* = \underset{T,\Theta,\Phi}{argmin}\, Loss(\hat{T}, \hat{I}, \hat{D}, \hat{\Phi}, \hat{\mathbf{F}} | I, D, \mathbf{F}), \tag{10}$$

where $T^*, \Theta^*, \Phi^*$ are optimized parameters, respectively.

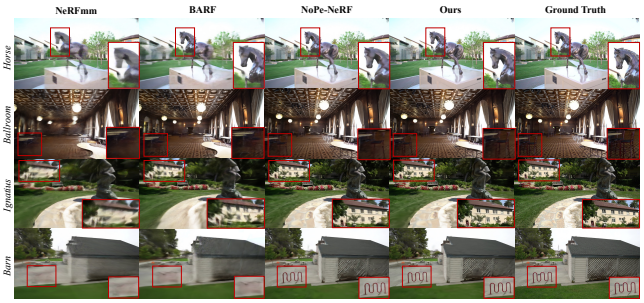



**Fig. 4**: **Qualitative results of novel view synthesis on Tanks and Temples dataset [13].**

## 3. EXPERIMENTS

### 3.1. Dataset

Our experiment is conducted on the Tanks and Temples [13] dataset, which contains challenging realistic indoor and outdoor environments with large camera movements. Following [10], we select 8 scenes to evaluate pose accuracy and novel view synthesis quality, with all images set to $960 \times 540$. The ground truth of camera parameters is estimated by COLMAP [2]. For a fair comparison, we select every 8-th image in the sequences for evaluation as defined in [1].

### 3.2. Implementation Details

Our model is built on NeRFmm [4]. We integrate the depth loss and the learnable mono-depth distortion parameter from [10] into NeRFmm. For extrinsics estimation, we initialize the translation of each image as a zero vector and the rotation matrix as an identity matrix. Camera rotations are optimized in axis-angle $\phi_i \in \mathfrak{so}(3)$. In each iteration, our proposed APAS strategy selects 1024 rays from an input image, and uniformly samples 128 points along each ray within the range of $(0.1, 10)$. The model is trained for 10k epochs per scene on an NVIDIA Tesla V100 GPU with two Adam optimizers for NeRF and other parameters. The initial learning rates of them are set to 0.001 and 0.0005, respectively.

### 3.3. Results on Tanks and Temples Dataset

We compare our method with state-of-the-art pose-free NeRF baselines, including NeRFmm, BARF, and NoPe-NeRF. Before the evaluation of each method, we follow NeRFmm to obtain the initial test poses by minimizing the photometric loss while keeping NeRF frozen. Qualitative and quantitative results of novel view synthesis are shown in Fig.4 and Tab.1, where the image quality metrics PSNR, SSIM, and LPIPS [23] are reported. The quantitative results of pose estimation are presented in Tab.2, we report the relative translation error ($\text{RPE}_t$), relative rotation error ($\text{RPE}_r$), and absolute trajectory error (ATE). Our method outperformed other baselines by a large margin in both NVS and pose estimation.

### 3.4. Ablation Study

Tab.3 shows the ablation experiments of our methods. We observe that our proposed flow-depth consistency guidance is the core contributor to the improvement of pose estimations. When the flow-depth consistency is not considered, the pose accuracy and NVS quality degrade significantly. This is due to the absence of direct guidance from the pose optimization gradient, making it more difficult for pose-free NeRF to optimize effectively and leading to pose-geometry ambiguity. When disabling the APAS strategy, as mentioned in Sec.2.3, the rays sampled by random sampling may provide less effective supervision compared to APAS, resulting in lower performance in pose accuracy and synthesis quality. With the effect of a collaboration of all introduced modules, our proposed method outperforms the baseline model and achieves state-of-the-art results.

**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 |

**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] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\text{RPE}_t \downarrow$ | $\text{RPE}_r \downarrow$ | ATE↓ | $\text{RPE}_t \downarrow$ | $\text{RPE}_r \downarrow$ | ATE↓ | $\text{RPE}_t \downarrow$ | $\text{RPE}_r \downarrow$ | ATE↓ | $\text{RPE}_t \downarrow$ | $\text{RPE}_r \downarrow$ | 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 |

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

| Methods | NVS | | | Pose Est. | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | $\text{RPE}_t \downarrow$ | $\text{RPE}_r \downarrow$ | 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 |

## 4. CONCLUSIONS

In this paper, we introduce FDC-NeRF that incorporates the flow-depth consistency guidance and Adaptive Pose-Aware Sampling (APAS) to jointly estimate camera poses and NeRF in scenes with large camera movement and without pose prior. The flow-depth consistency guidance leverages the consistency between 2D optical flow and 3D virtual flow to provide effective directions for pose optimization. We further propose the novel APAS to select pose-aware interest points and adaptively adjust the region to the entire image for better joint training. Qualitative and quantitative results on the Tanks and Temples dataset demonstrate our state-of-the-art performance on both novel view synthesis and pose estimation.

# References


[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*. Springer, 2020, pp. 405–421.

[2] Johannes L Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin, "inerf: Inverting neural radiance fields for pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1323–1330.

[4] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu, "NeRF−−: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[5] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey, "Barf: Bundle-adjusting neural radiance fields," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 5721–5731.

[6] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park, "Self-calibrating neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5846–5854.

[7] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey, "Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 264–280.

[8] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool, "Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction," in *33rd British Machine Vision Conference*, 2022.

[9] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu, "Gnerf: Gan-based neural radiance field without posed camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6351–6361.

[10] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu, "Nope-nerf: Optimising neural radiance field with no pose prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4160–4169.

[11] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv preprint arXiv:2010.07492*, 2020.

[12] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[13] A. Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, "Tanks and temples: benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, 2017.

[14] James T Kajiya and Brian P Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[15] Nelson Max, "Optical models for direct volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 2, pp. 99–108, 1995.

[16] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou, "Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5610–5619.

[17] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12179–12188.

[18] Huachen Gao, Xiaoyu Liu, Meixia Qu, and Shijie Huang, "Pdanet: Self-supervised monocular depth estimation using perceptual and data augmentation consistency," *Applied Sciences*, vol. 11, no. 12, pp. 5383, 2021.

[19] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8121–8130.

[20] Simon Meister, Junhwa Hur, and Stefan Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[21] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[22] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.