Multimodal LiDAR-Camera Novel View Synthesis with Unified Pose-free Neural Fields

Weiyi Xue[1,2], **Fan Lu**[1†], **Yunwei Zhu**[1], **Zehan Zheng**[1],
Haiyun Wei[1], **Sanqing Qu**[1], **Jiangtong Li**[1], **Ya Wu**[3], **Guang Chen**[1,2†]

{xwy, lufan, 2432040, zhengzehan, 2311399, sanqingqu, jiangtongli, guangchen}@tongji.edu.cn,
qpo144@163.com,
[1] Tongji University, [2] Shanghai Innovation Institute,
[3] CNNC Equipment Technology Research (Shanghai) Co., Ltd.
[†] Corresponding author

Abstract

Pose-free Neural Radiance Field (NeRF) aims at novel view synthesis (NVS) without relying on accurate poses, exhibiting significant practical value. Image and LiDAR point cloud are two pivotal modalities in autonomous driving scenarios. While demonstrating impressive performance, single-modality pose-free NeRFs often suffer from local optima due to the limited geometric information provided by dense image textures or the sparse, textureless nature of point clouds. Although prior methods have explored the complementary strengths of both modalities, they have only leveraged inherently sparse point clouds for discrete, non-pixel-wise depth supervision, and are limited to NVS of images. As a result, a **M**ultimodal **U**nified **P**ose-free framework remains notably absent. In light of this, we propose **MUP**, a pose-free framework for LiDAR-Camera joint NVS in large-scale scenes. This unified framework enables continuous depth supervision for image reconstruction using LiDAR-Fields rather than discrete point clouds. By leveraging multimodal inputs, pose optimization receives gradients from the rendering loss of point cloud geometry and image texture, thereby alleviating the issue of local optima commonly encountered in single-modality pose-free tasks. Moreover, to further guide pose optimization of NeRF, we propose a multimodal geometric optimizer that leverages geometric relations from point clouds and photometric regularization from adjacent image frames. Besides, to alleviate the domain gap between modalities, we propose a multimodal-specific coarse-to-fine training approach for unified, compact reconstruction. Extensive experiments on KITTI-360 and NuScenes datasets demonstrate MUP's superiority in accomplishing geometry-aware, modality-consistent, and pose-free 3D reconstruction.

1 Introduction

Neural Radiance Fields (NeRFs) [20] have made substantial strides in novel view synthesis (NVS) for images and LiDAR point clouds [37, 54, 56, 12], with promising applications in autonomous driving scenarios [49, 40, 46, 47]. Recent developments have transitioned towards a pose-free paradigm [16, 10, 25, 3], facilitating reconstruction while accurately estimating sensor poses. This approach reduces dependence on time-consuming structure-from-motion algorithms like COLMAP [33] and on unreliable point cloud registration methods such as ICP [2, 29, 32, 28], both of which are susceptible to failure in wide-baseline scenarios.

However, existing pose-free NeRFs have largely concentrated on single modalities, particularly on images. Nevertheless, due to the lack of geometric consistency, relying solely on rich texture without

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

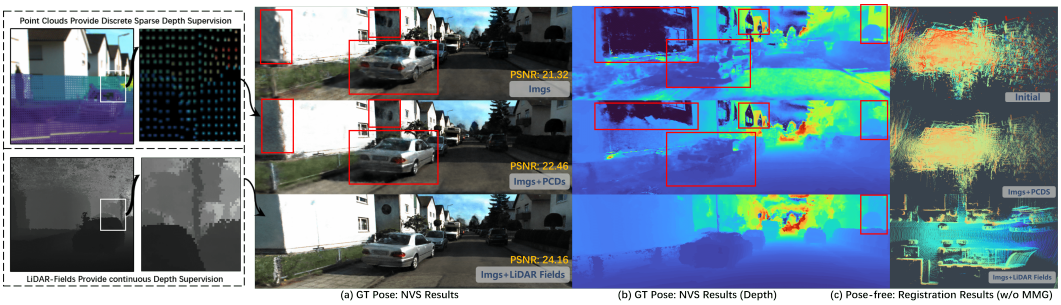

Figure 1: **NVS results w/ and w/o accurate poses.** Compared to continuous LiDAR-Camera Fields, projecting LiDAR point clouds onto images as discrete depth priors fails to provide continuous, pixel-wise supervision. Multimodal NeRFs (without MMG) leverage continuous LiDAR-Fields to constrain geometric consistency and optimize pose, aiding both reconstruction and pose optimization.

geometry often leads to suboptimal results in large-scale scenes [19]. The assistance of Pseudo or real point clouds [3] can help address this issue. For instance, directly propagating depth loss to optimize geometric consistency [48] and leveraging depth for correspondence establishment and reprojection refinement [39, 3] can both contribute to pose optimization. Nonetheless, when performing pose-free reconstruction and projecting discrete point clouds onto images for depth supervision, only a sparse set of pixels contains depth information, underutilizing point cloud geometry. As shown in Fig. 1, LiDAR point clouds fail to provide pixel-wise supervision due to their inherent sparsity and discrete sampling. Comparing Fig. 1 reveals that supervision with discrete point clouds limits registration accuracy(c) and reconstruction quality(a), whereas continuous LiDAR-Field supervision excels in both. Moreover, Nope-NeRF [3] employs pixel-wise depth estimation to generate pseudo point clouds but suffers from scale ambiguity and limited accuracy. Consequently, even with point cloud [3, 36], image-based pose-free NeRFs [16, 19, 3] remain challenging to apply in large-scale scenes and are unable to perform point cloud NVS. Conversely, instead of utilizing point clouds as additional discrete supervision, recent advancements [37] have demonstrated that leveraging point clouds alone enables the reconstruction of continuous Neural LiDAR Fields. The LiDAR-based framework [56, 37, 12] facilitates highly accurate and continuous geometric reconstructions. Among these efforts, GeoNLF [48] extends LiDAR-Field to pose-free reconstruction task. Nonetheless, the inherent lack of texture information in point clouds, along with the sparse sampling and indistinct foreground-background boundaries in the range map, continues to constrain the performance.

Regarding the challenges encountered in the aforementioned single-modality approach, Continuous Neural LiDAR Fields can provide pixel-wise depth supervision for images and directly propagate gradients to pose estimation, providing continuous geometric constraints. In turn, images offer rich textures and clear boundaries, which enhance the registration accuracy of sparse point clouds. Consequently, reconstructing both point clouds and images as continuous neural fields allows them to effectively complement each other in pose-free scenarios. Nevertheless, prior research [38] has faced challenges due to the significant domain gap and uncoordinated convergence problems [27, 43, 35] between these modalities. Therefore, Alignmif [38] employs independent hash-grids for each modality. However, in pose-free, ill-conditioned optimization, jointly optimizing the two distinct hash grids and poses is infeasible and yields suboptimal performance compared to the single-modality model. The large discrepancy between two feature spaces leads to inconsistent gradients when propagated to poses, causing [38] to fail to converge.

Consequently, in pursuit of effectively integrating the two modalities for unified pose-free reconstruction, we introduce MUP—a framework that facilitates the simultaneous reconstruction of both point clouds and images via a unified neural field. Specifically, to mitigate local minima issues in single-modality approaches, we propagate the gradient of multi-modal rendering loss to poses with varying emphasis at different optimization stages. Additionally, the MultiModal Geometric optimizer (MMG) guides pose optimization by leveraging geometric relations between multiview point clouds and incorporating point-to-image error as a regularization term. To alleviate modality conflicts [38] and address the uncoordinated convergence problem, we introduce a multimodal-specific coarse-to-fine training approach [16], facilitating the utilization of a singular hash grid for compact reconstruction. Moreover, to enhance color-depth consistency, we introduce a consistency constraint by projecting image pixels onto adjacent frames using depth derived from NeRF. Therefore, MUP is capable of achieving geometry-aware, modality-consistent, and pose-free reconstruction in large-scale scenarios.

We evaluate our method across diverse scenarios using the KITTI-360 [15] and NuScenes [4] autonomous driving datasets. Comprehensive experiments demonstrate that MUP significantly outperforms prior state-of-the-art techniques and single-modality approaches by a large margin in both registration and NVS.

In summary, our primary contributions can be delineated as follows: (1) We propose MUP, a unified pose-free framework that combines the advantages of two modalities for pose estimation and multimodal NVS in large-scale scenes, efficiently leveraging a compact neural representation without the need for accurate poses. (2) We introduce a multimodal-specific training approach, integrated with the MMG module and consistency constraint, to facilitate modality-consistent, pose-free, and geometry-aware reconstruction. (3) We demonstrate the effectiveness of our method quantitatively and qualitatively through extensive experiments conducted on multiple datasets and scenes.

2 Related Work

NeRF for Single-Modality NVS. NeRF [20] and related works have achieved substantial progress in novel view synthesis. Diverse neural representations [21, 1, 5, 6, 11], techniques [22, 23, 42, 53], and generalization methods [7] for NeRF have been introduced to enhance its performance. Some methods incorporate depth priors [9, 31, 52] or point clouds as auxiliary data to ensure multi-view geometric consistency. However, relying solely on sparse depth supervision from point clouds fails to fully exploit their potential in expressing geometry. Consequently, researchers have extended the NeRFs to generate novel views from LiDAR [37, 12, 56, 54, 48], treating point clouds as range images. Nevertheless, sparse point clouds are notably deficient in dense texture information. Accordingly, we aim to leverage the complementary characteristics of both modalities, advancing a unified multimodal NeRF framework.

Multimodal Joint Learning in NeRF. NeRF framework facilitates the integration of a wide range of attributes into the volumetric rendering pipeline, including color [21], depth, intensity, and semantic labels [54]. Recent advancements [3, 9, 14, 30, 41] exploit point clouds to provide depth priors but fail to offer pixel-wise supervision. Neural sensor simulator Unisim [50] performs multimodal NVS via implicit fusion. However, all these methods rely on accurate poses. Very recently, Alignmif [38] has proposed using distinct hash grids for separate modality reconstruction followed by fusion. However, its intricate structure with two hash grids fails to converge in pose-free optimization and is highly computationally expensive. Our approach employs a more compact representation and introduces a novel strategy to achieve pose-free, multimodal reconstruction.

NeRF with Pose Optimization. Since iNeRF [51] and subsequent works [17, 8] demonstrated that NeRF can optimize the poses of new viewpoint images based on trained radiance fields, a series of approaches have aimed to reduce NeRF's reliance on highly accurate poses. NeRFmm [45] and SCNeRF [34] extend the method to intrinsic parameter estimation. BARF [16, 10] employs a coarse-to-fine reconstruction scheme that gradually learns positional encodings, demonstrating notable efficacy. Additionally, several studies have expanded BARF to tackle more challenging scenarios, such as sparse input [39], dynamic scenes [56, 18], and generalizable NeRF [7]. The coarse-to-fine training method has been particularly inspiring for our work. However, these pioneering efforts primarily target indoor or object-level scenes. Increasingly, pose-free methods have enhanced robustness by incorporating priors. In particular, [3, 39] use monocular depth or correspondence priors for constraints. Recently, [48] proposed a LiDAR-only pose-free framework. Nonetheless, the sparse nature of point clouds, coupled with the absence of texture information, ray-drop characteristics, and inherent noise, imposes limitations on registration accuracy. Moreover, all of the aforementioned methods are designed for a single modality. In the context of autonomous driving, they fail to fully exploit both the geometric information from point clouds and the texture information from images. As a result, a unified multimodal, pose-free framework remains absent.

3 Preliminaries

Pose-free NeRF for Images and Point Clouds. NeRF represents a 3D scene implicitly by encoding the density σ along with additional data features such as color and intensity of the scene using an implicit neural function $F_\Theta(\boldsymbol{x}, \boldsymbol{d})$, where $\boldsymbol{x}$ is the 3D coordinates and $\boldsymbol{d}$ is the view direction. NeRF pipeline is compatible with both point clouds and images. For point clouds, it converts LiDAR point

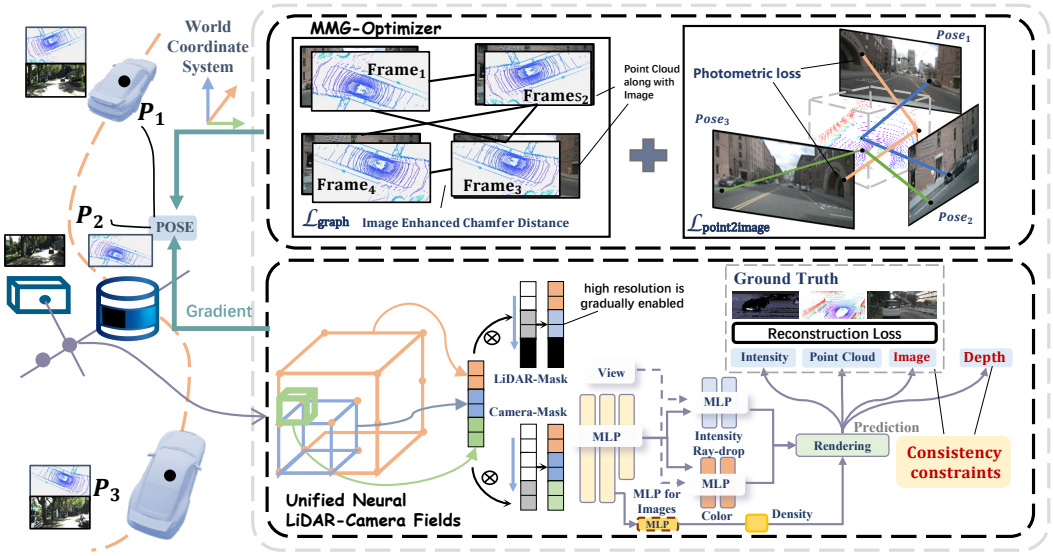

Figure 2: **Overview of our proposed MUP.** MUP derives pose gradients through both implicit global optimization from the Unified Neural LiDAR-Camera Fields and our explicit MMG-optimizer, both of which effectively leverage complementary multimodal information. Besides, we integrate Unified Neural LiDAR-Camera Fields with a multimodal-specific coarse-to-fine training strategy, along with consistency constraint to achieve geometry-aware, modality-consistent and pose-free reconstruction.

clouds into a range image, then casts a ray with a direction $\boldsymbol{d}$ determined by the azimuth angle θ and elevation angle ϕ under the polar coordinate system: $\mathbf{d} = (\cos\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\phi)^T$. When performing NVS, NeRF employs volume rendering techniques to accumulate densities and the pixel depth value $\hat{\mathcal{D}}$ along sampled rays. Using the same approach, NeRF predicts the color $\mathcal{C}$ for images or the intensity $\mathcal{S}$ and the ray-drop $\mathcal{R}$ for point clouds.

Traditional NeRF relies heavily on accurate sensor poses and reconstruction accuracy can be significantly compromised with imprecise poses. Pose-free NeRF is introduced to solve this issue by treating sensor poses $P = \{P_s | s = 0, 1...N-1\}$ as optimizable parameters. Hence, the simultaneous update via gradient descent of P and Θ can be achieved by minimizing $\mathcal{L} = \sum_{i=0}^{N} \|\hat{I}_i - I_i\|_2^2$ between the rendered and ground truth image or range image $\hat{I}, I$:

$$\Theta^*, \mathcal{P}^* = \arg\min_{\Theta,\mathcal{P}} \mathcal{L}(\hat{\mathcal{I}}, \hat{\mathcal{P}} \mid \mathcal{I}). \tag{1}$$

Problem Formulation. In large-scale autonomous driving scenarios, given time-synchronized sequences of cameras and LiDAR data, denoted as $\mathcal{I} = \{I_s | s = 0, 1, \ldots, N-1\}$ and $\mathcal{Q} = \{Q_s | s = 0, 1, \ldots, N-1\}$, the objective of MUP is to reconstruct the scene as a continuous implicit representation based on a unified neural field. MUP is capable of performing NVS for both modalities, while also simultaneously recovering the vehicle poses $P = \{P_s | s = 0, 1, \ldots, N-1\}$, which enables the global alignment of both images and point clouds. The relative poses of all sensors with respect to the vehicle are assumed to be known.

Pose Representation. Following [3], pose is modeled as a rotation $\boldsymbol{R} \in SO(3)$ and a translation $\boldsymbol{t} \in \mathbb{R}^3$. This formulation allows for independent updates of the translation at the origin and the rotation around the origin. Rotation updates are computed in the Lie algebra of the special orthogonal group in three dimensions, $\boldsymbol{\phi} \in \mathfrak{so}(3)$, while translations are updated in $\mathbb{R}^3$. Specifically, the updates are expressed as $\boldsymbol{\phi}' = \boldsymbol{\phi} + \Delta\boldsymbol{\phi}$ and $\boldsymbol{t}' = \boldsymbol{t} + \Delta\boldsymbol{t}$. Here, $\boldsymbol{\phi}$ satisfies $\boldsymbol{R} = \sum_{n=0}^{\infty} \frac{1}{n!}(\boldsymbol{\phi}^\wedge)^n$, where $\boldsymbol{\phi}^\wedge$ denotes the skew-symmetric matrix representation of $\boldsymbol{\phi}$.

Challenges in Multimodal NeRF. Different modalities exhibit varying representations and converge at different rates [27, 43, 35]. Specifically, point clouds often converge faster than images due to their sparsity and direct geometric supervision. This disparity causes uncoordinated convergence, with the model shifting focus to images once the point cloud error is sufficiently small.

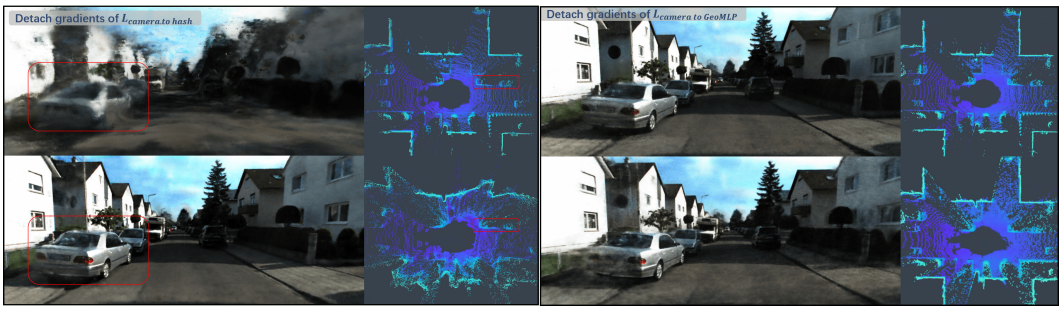

Figure 3: **Modality fusion in Hash-grids and geo-MLP.** We truncate the gradients of each modality separately in hash grids and geo-MLP. The results show that feature fusion across modalities primarily occurs in the hash grids rather than the geo-MLP.

4 Methodology

As shown in Fig. 2, our framework is divided into two main modules: a Unified neural fields that implicitly refines neural network and poses, and the Multimodal Geometric Optimizer (MMG) that explicitly optimizes poses. Both modules leverage geometric and texture information for registration and are executed alternately. In the following sections, we first present our LiDAR-Camera Fields in Section 4.1, integrating unified neural fields with a multimodal-specific coarse-to-fine training strategy for reconstruction and global implicit pose optimization. Then, we introduce our MMG module in Section 4.2, which provides explicit geometric guidance to avoid local optima. Finally, we present the proposed consistency constraint and the overall optimization pipeline in Section 4.3.

4.1 Unified Neural LiDAR-Camera Fields

To address the issue of unbalanced convergence speeds mentioned in Section 3, we design a Unified Neural LiDAR-Camera representation. Firstly, we introduce the Unified Neural Fields and identify that modality fusion occurs within the hash grids. Based on this observation, we propose a multimodal training method for optimizing the hash grid, which also stabilizes pose optimization and mitigates modality conflicts. Finally, a comprehensive analysis and discussion are provided.

Unified Neural Fields. Initially, we adopt i-NGP [21] as the base framework, leveraging multi-resolution hash grids to encapsulate the features, while a geometry-MLP (geo-MLP) is utilized to derive the density. In MUP, both the hash grids and the geo-MLP are shared across the modalities. For the image modality, *we use a lightweight MLP to refine the geo-MLP output, helping reduce modality conflicts.* To explore how modality features are fused, we independently truncate the gradients of reconstruction loss L_{Camera} and L_{LiDAR} to *hash grids* and *geo-MLP*. For hash grids, results indicate that truncating one modality prevents the Multimodal NeRF from learning the corresponding features. As shown in the upper images of Fig. 3, when the image gradient is truncated, novel views lose texture and resemble a colored point cloud projection, whereas truncating the point cloud gradient results in inaccurate geometry, resembling image-based pseudo point clouds. The same experiment on *geo-MLP* reveals slight performance degradation, suggesting feature fusion primarily occurs in the hash grids. *Thus, effectively controlling hash grid learning across modalities is crucial.*

Multimodal-specific Coarse-to-fine Training. To this end, we draw inspiration from the coarse-to-fine (C2F) strategy, which is widely used in pose-free NeRFs [48, 16, 10]. We extend this approach to a multimodal pose-free NeRF by adopting modality-specific C2F strategies, which helps to balance the influence of each modality on the hash grid. Specifically, we progressively activate shared hash grids from low to high resolution, employing distinct activation speeds and initiation points for each modality, as described in Eq. (2).

$$\gamma'^{\text{LiDAR}}_{L} = w_L(\alpha_{\text{LiDAR}})\gamma^{\text{LiDAR}}_{L}, \gamma'^{\text{Camera}}_{L} = w_L(\alpha_{\text{Camera}})\gamma^{\text{Camera}}_{L}, \tag{2}$$

where γ'_L denotes the encoding of the L-th layer hash-grid, w_L is the coarse-to-fine mask, and it can be any monotonic increasing function with a domain of $[L-1, L]$ and a range of $[0, 1]$, such as $w_L(\alpha) = \text{clip}(\alpha - L + 1, 0, 1)$ or $w_L(\alpha) = (1 - \cos(\text{clip}(\alpha - L + 1, 0, 1)\pi))/2$ in most pose-free methods [16, 48, 10]. $\alpha \in [0, L]$ is a controllable parameter proportional to the optimization progress. Notably, α varies across modalities. For images, all low-resolution hash grids are initially activated, with higher-resolution grids progressively activated. For the point cloud, a similar coarse-to-fine approach is used but with slower activation starting from low-resolution grids. In our experiment, the α_{LiDAR} is adjusted between 6-16, while the α_{Camera} is adjusted between 12-16.

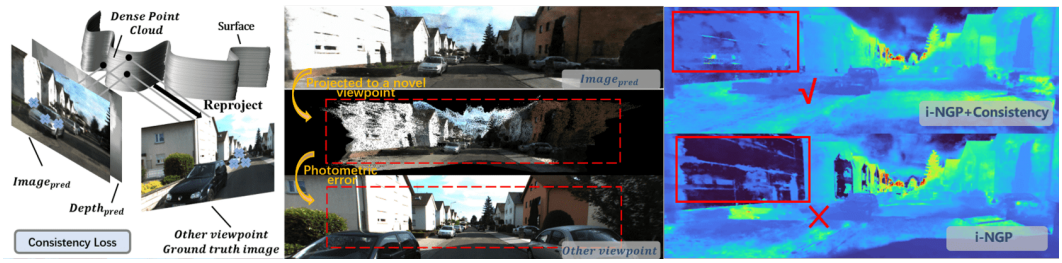

Figure 4: **Consistency constraint.** We project rendered images onto other frames by depth obtained from NeRF to compute the photometric error. It's particularly effective for textureless regions.

Implicit Pose Optimization. In the Unified NeRF training, gradients are also propagated to pose from reconstruction loss. However, in early optimization, geometric inaccuracies hinder texture-based pose refinement. In later stages, sparse point clouds limit registration accuracy, while images offer denser cues for alignment. Consequently, we adopt point cloud-based loss at the early stages and later employ image-based photometric loss to refine the poses. This is implemented by adjusting the learning rates of pose parameters across different modalities, as depicted in Eq. (3).

$$\mathbf{P}_{n+1} = \mathbf{P}_n - (1 - \mathtt{w})\mathtt{lr}G_{\text{LiDAR}} - \mathtt{w} \cdot \mathtt{lr}G_{\text{Camera}}, \tag{3}$$

where G is the gradient of the corresponding modality, $\mathbf{P}_n$ denotes the pose at the n-th iteration, $\mathtt{w}$ is a control variable increasing progressively during the training process, $\mathtt{lr}$ denotes the learning rate.

Discussion. The C2F strategy is widely used in pose-free NeRFs [48, 16, 10]. During ill-conditioned optimization, minor perturbations in pose can lead to significant deviations in NeRF, potentially driving it towards a local minimum. The C2F strategy alleviates this issue by blocking partial gradient propagation, thereby mitigating the impact of such perturbations. The Jacobian of γ'_L thus becomes:

$$\frac{\partial \gamma'_L(\theta, \mathbf{x}; \alpha)}{\partial \theta} = w_L(\alpha)\frac{\partial \gamma_L(\theta, \mathbf{x})}{\partial \theta}, \frac{\partial \gamma'_L(\theta, \mathbf{x}; \alpha)}{\partial \mathbf{x}} = w_L(\alpha)\frac{\partial \gamma_L(\theta, \mathbf{x})}{\partial \mathbf{x}} \tag{4}$$

where θ denotes the parameters of the hash grids, and the point $\mathbf{x}$ is associated with the pose. When $w_L(\alpha) = 0$, the contribution to the gradient from the L-th (and higher) resolution component is nullified. As shown in Eq. (4), the optimization of both hash-grid and pose follows a coarse-to-fine strategy. *In the early stages of optimization, only the gradients from the coarse resolution of the hash grids contribute to pose optimization, while the finer resolutions further refine the pose.*

Furthermore, our method utilizes a single shared hash grid and unified hash features across modalities. We adjusts the α for each modality, ensuring consistent convergence speeds and balanced loss across modalities. Moreover, the varying α values guide high-resolution hash grids to capture fine image textures, while the LiDAR field refines geometry and primarily supervises low-frequency geometric structures, mitigating cross-modal conflicts. In summary, our method ensures synchronous convergence and stable optimization, while also mitigating modality conflicts, by employing a distinct, compact, and efficient hash grid structure.

4.2 Multimodal Geometric Optimizer

Our MMG module leverages Image-enhanced Chamfer Distance combined with point-to-image regularization. Unlike implicit optimization through NeRF, MMG explicitly optimizes poses by leveraging both geometry and textures.

Explicit Pose Optimization. The closest-point correspondences between two partially overlapping point clouds establish the most direct geometric relationship, which can be leveraged effectively for registration like ICP [2]. Chamfer Distance (CD) is a well-established loss derived from point correspondences and can be computed as Eq. (5):

$$\mathbf{CD}_{(P,Q)} = \sum_{\mathbf{p}_i \in \mathbf{P}} w_i \min_{\mathbf{q}_i \in \mathbf{Q}} \|\mathbf{T_P} p_i - \mathbf{T_Q} q_i\|_2^2 + \sum_{\mathbf{q}_i \in \mathbf{Q}} w_i \min_{\mathbf{p}_i \in \mathbf{P}} \| \mathbf{T_Q} q_i - \mathbf{T_P} p_i\|_2^2, \tag{5}$$

where q, p in point cloud $\mathbf{Q}, \mathbf{P}$ are homogeneous coordinates. $\mathbf{T}_P, \mathbf{T}_Q$ represent the transformation matrix to the world coordinate system. Additionally, we define w_i to represent the weight of each correspondence. Our MMG module directly computes the inter-frame CD and propagates the gradient

Table 1: **Quantitative comparison of NVS in pose-free setting.** We conduct experiments under the pose-free setup. The estimated trajectory is aligned with the ground truth using Sim(3) for image-based methods. $\mathcal{PF}$: Pose-Free, $\mathcal{RR}$: Reconstruction after Registration, $\mathcal{I}$: Image-synthesizable, $\mathcal{PI}$: Image and Point cloud-synthesizable.

Methods	Type	LiDAR Metrics			Image Metrics			Pose Metrics		
		CD↓	F-score↑	MAE_I↓	PSNR↑	SSIM↑	LPIPS↓	RPE_t (cm)↓	RPE_r(deg) ↓	ATE(m)↓
Experiments on KITTI-360 [15]										
Colored-ICP [24, 38]	$\mathcal{RR}$./ $\mathcal{PI}$	0.492	0.787	0.149	20.92	0.698	0.459	25.383	0.899	1.624
Nope-NeRF [3, 57]	$\mathcal{PF}$./ $\mathcal{I}$	-	-	-	19.82	0.337	0.592	83.223	14.412	0.653
BA-Alignmif [38, 10]	$\mathcal{PF}$./ $\mathcal{PI}$	0.641	0.722	0.116	19.12	0.641	0.439	36.179	0.498	0.410
MUP(Ours)	$\mathcal{PF}$./ $\mathcal{PI}$	0.079	0.942	0.096	23.46	0.759	0.287	1.471	0.025	0.187
Experiments on NuScenes [4]										
Colored-ICP [24, 38]	$\mathcal{RR}$./ $\mathcal{PI}$	0.930	0.599	0.047	19.21	0.438	0.644	14.380	0.599	1.170
Nope-NeRF [3, 57]	$\mathcal{PF}$./ $\mathcal{I}$	-	-	-	18.01	0.341	0.670	129.899	12.399	0.718
BA-Alignmif [38, 10]	$\mathcal{PF}$./ $\mathcal{PI}$	1.695	0.603	0.044	18.73	0.621	0.619	182.391	0.377	4.266
MUP(Ours)	$\mathcal{PF}$./ $\mathcal{PI}$	0.810	0.656	0.042	20.83	0.699	0.585	4.058	0.101	0.176

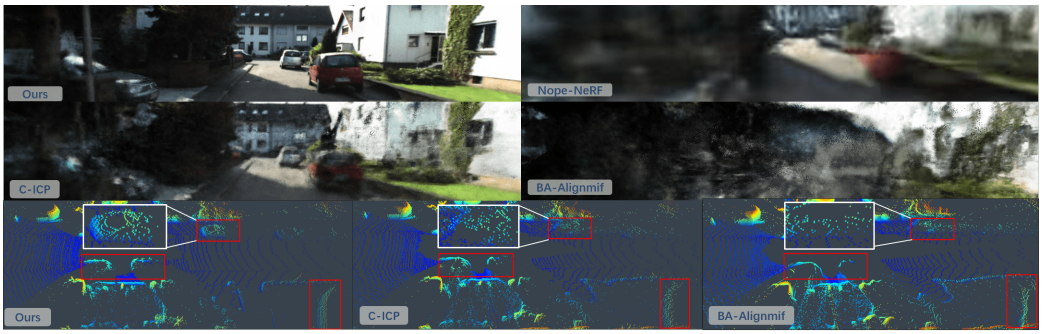

Figure 5: **Qualitative comparison of NVS.** We compared MUP with pose-free and registration-first methods. Nope-NeRF and Colored-ICP-assisted fail due to the large-scale scene. BA-Alignmif struggles to converge. All baselines fail entirely on certain sequences.

to the poses, guiding the optimization of the poses during the ill-conditioned optimization. However, CD overlooks the partial overlap of point clouds, merely minimizing CD does not necessarily improve pose accuracy. Leveraging the multimodal input, we exploit images to alleviate the impact of non-overlapping regions. Specifically, we project point clouds onto time-synchronized images to derive color information. During the calculation of CD, we re-weight the relevant correspondences by incorporating photometric information.

$$w_i = \begin{cases} 1 & \|C\langle\mathcal{F}(p_i)\rangle - C\langle\mathcal{F}(q_i)\rangle\|_1 \leq m \\ 0 & \|C\langle\mathcal{F}(p_i)\rangle - C\langle\mathcal{F}(q_i)\rangle\|_1 > m, \end{cases} \tag{6}$$

where the $C\langle\cdot\rangle$ denotes the color obtained by sampling from the image. $\mathcal{F}(x)$ denotes the projection function that maps a 3D point x onto a 2D image. Additionally, following [48], we further construct a multi-frame graph $(\mathcal{S}, \mathcal{E})$, where vertex $\mathcal{S}$ represents a frame of point cloud along with the time-synchronized image, each edge $\mathcal{E}$ represents the CD. Finally, the Image-enhanced CD is computed as $\mathcal{L}_{\text{ICD}} = \sum_{(i,j)\in\mathcal{E}} \mathbf{CD}_{(i,j)}$.

Point-to-Image Regularization. As shown in Fig. 2 (top right), we project point clouds onto images from adjacent frames to establish pixel correspondences. Based on these correspondences, we introduce a point-to-image error using photometric loss, which serves as a regularization term.

$$\mathcal{L}_{\text{Point2Image}} = \sum_{ij} \|C_i\langle\mathcal{F}_i(P_i)\rangle - C_j\langle\mathcal{F}_j(T_j T_i^{-1} P_i)\rangle\|_2^2, \tag{7}$$

where P denotes the point cloud. Finally, the overall optimization loss for MMG is defined as:

$$\mathcal{L}_{\text{MMG}} = \mathcal{L}_{\text{IRCD}} + \mathcal{L}_{\text{Point2Image}}. \tag{8}$$

4.3 Overall Optimization Pipeline

Consistency Constraint. Due to the differing Fields of View (FoV) between LiDAR and cameras, plain LiDAR-Fields provide incomplete and limited depth supervision. Therefore, we further introduce a geometric consistency constraint, which leverages reprojection error to constrain regions outside the LiDAR's FoV. Specifically, we extract dense point clouds from rendered depth maps z and project them onto adjacent frame images to compute photometric errors. It is effective for large textureless regions, enforcing geometric-color consistency, and is calculated as Eq. (9):

$$\mathcal{L}_{\text{cons}} = \sum_{(i,j)} \| C_i\langle p \rangle) - C_j\langle \mathcal{F}_j(T_j T_i^{-1} \mathcal{F}^{-1}(z,p)) \rangle \|_2^2, \tag{9}$$

where $\mathcal{F}^{-1}(z,p)$ denotes the back-projection, mapping a pixel p to a 3D point using depth z.

Optimization. To optimize MUP, the total reconstruction loss is formulated as a weighted sum of intensity loss $\mathcal{L}_S$, ray-drop loss $\mathcal{L}_R$, LiDAR range image loss $\mathcal{L}_D$, photometric loss $\mathcal{L}_{rgb}$ from the image and consistency loss $\mathcal{L}_{\text{cons}}$:

$$\mathcal{L}_{\text{D}}(\mathbf{r}) = \sum_{\mathbf{r}\in R} \left\| \hat{D}(\mathbf{r}) - D(\mathbf{r}) \right\|_1, \mathcal{L}_{\text{rgb}}(\mathbf{r}) = \sum_{\mathbf{r}\in R} \left\| \hat{I}(\mathbf{r}) - I(\mathbf{r}) \right\|_2^2 \tag{10}$$

$$\mathcal{L}_{\text{S}}(\mathbf{r}) = \sum_{\mathbf{r}\in R} \left\| \hat{S}(\mathbf{r}) - S(\mathbf{r}) \right\|_2^2, \mathcal{L}_{\text{R}}(\mathbf{r}) = \sum_{\mathbf{r}\in R} \left\| \hat{R}(\mathbf{r}) - R(\mathbf{r}) \right\|_2^2 \tag{11}$$

$$\mathcal{L} = \lambda_\alpha \mathcal{L}_{\text{D}} + \lambda_\beta \mathcal{L}_{\text{rgb}} + \lambda_\gamma \mathcal{L}_{\text{S}} + \lambda_\eta \mathcal{L}_{\text{R}} + \lambda_{\text{r}} \mathcal{L}_{\text{cons}} \tag{12}$$

5 Experiment

5.1 Experimental Setup

Datasets and Experimental Settings. We conducted experiments on two public autonomous driving datasets: NuScenes [4] and KITTI-360 [15] dataset, each with five representative time-synchronized LiDAR point cloud and image sequences. For the NuScenes dataset, it includes six cameras and a LiDAR sensor, with keyframes that are typically used, which are time-synchronized based on timestamps. Following [48] we selected 33 consecutive frames from keyframes as a single scene. KITTI-360 has three cameras and a LiDAR, where each frame's point cloud and image are time-aligned. Following [37, 56, 38], we use the standard KITTI-360 dataset, all images and point clouds are time-synchronized. For both datasets, only the front-facing single camera was utilized. Following [48, 16], we perturbed poses of car with additive noise corresponding to a standard deviation of $20\,\text{deg}$ in rotation and $3m$ in translation. The relative poses of all sensors with respect to the vehicle are assumed to be provided.

Metrics. We evaluate our method for pose estimation and NVS. For pose estimation, we follow [3], employing standard odometry metrics: Absolute Trajectory Error (ATE) and Relative Pose Error (RPE), with rotational (RPE_r) and translational (RPE_t) components. Following [37, 56] for point cloud NVS, we adopt CD to assess 3D geometric errors and the F-score with a 5 cm threshold. We also compute mean absolute error (MAE) for intensity in projected range images. Besides, we follow the approach in [3, 38], employing PSNR, LPIPS [55], and SSIM [44] for image NVS.

Implementation Details. All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU. 768 points were uniformly sampled along each ray for two modalities. MUP optimization was implemented in PyTorch [26] using the Adam optimizer [13]. The learning rates were set as follows: 1×10^{-2}, decaying to 1×10^{-4} for NeRF; 1×10^{-3}, decaying to 1×10^{-5} for translation; and 5×10^{-3}, decaying to 5×10^{-5} for rotation. The weighting coefficients for each loss term are defined as: $\lambda_\alpha = \lambda_\beta = 1000, \quad \lambda_\gamma = 10, \quad \lambda_\eta = 2.5, \quad \lambda_r = 150$. Besides, all sequences are trained for 60K iterations in the pose-free setting and 30K iterations when ground truth poses are available. Additionally, after every m_1 epoch of Unified NeRF training, we proceed with m_2 epochs of pure geometric optimization, where the ratio m_2/m_1 decreases from 10 to 1. Both α values are tuned so that the coarse-to-fine strategy is applied during the training progress between 0 and 0.3. In the training process, `w` increases from 0 to 1.

Methods(Pose - free)	LiDAR		Image	Pose Metrics		
	CD↓	F-score↑	PSNR↑	RPE_t ↓	RPE_r ↓	ATE↓
w/o MMG	0.592	0.731	19.27	26.201	0.433	0.805
w/o $\mathcal{P}2\mathcal{IR}$	0.083	0.936	23.35	1.533	0.041	0.205
w/o Image	0.089	0.937	-	1.668	0.062	0.224
w/o MSC2F	0.113	0.932	22.06	1.542	0.058	0.256
MUP(Ours)	0.079	0.942	23.46	1.471	0.025	0.187

Methods(GT - pose)	LiDAR Metrics			Image Metrics		
	CD↓	F-score↑	MAE_I ↓	PSNR↑	SSIM↑	LPIPS↓
w/o Cons	0.092	0.931	0.096	23.66	0.793	0.227
w/o MSC2F	0.113	0.923	0.102	23.24	0.798	0.230
MUP(Ours)	0.080	0.945	0.089	24.29	0.812	0.211

Table 2: **Ablation studies on the MMG module and image modality under the pose-free setting(top).** MMG module plays a pivotal role in pose optimization. **Ablations of MSC2F and consistency constraint under GT-pose setting(bottom).** $\mathcal{P}2\mathcal{IR}$: Point2Image Regularization. MSC2F: Training Strategy.

Methods	LiDAR Metrics			Image Metrics		
	CD ↓	F-score ↑	MAE_I ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Experiments on KITTI - 360 [15], i-NGP: i-NGP w/ point cloud.						
i-NGP [21]	-	-	-	23.12	0.791	0.223
L-NeRF [37]	0.083	0.942	0.097	-	-	-
i-NGP [21]	-	-	-	23.23	0.794	0.220
MUP(Ours)	0.080	0.945	0.089	24.29	0.812	0.211
Experiments on NuScenes [4] , i-NGP: i-NGP w/ point cloud.						
i-NGP [21]	-	-	-	20.78	0.667	0.530
L-NeRF [37]	0.815	0.673	0.041	-	-	-
i-NGP [21]	-	-	-	20.92	0.682	0.564
MUP(Ours)	0.798	0.678	0.038	21.53	0.704	0.545

Table 3: **Quantitative comparison on NVS with GT-poses.** We conducted experiments under GT-poses to demonstrate the effectiveness of our method in modal fusion.

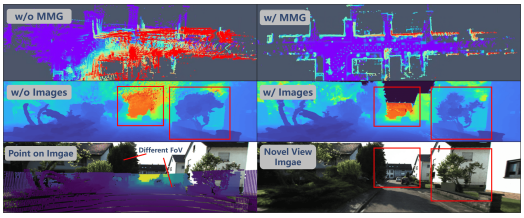

Figure 6: **Ablation in pose-free setting.** The first row illustrates registration results w/ and w/o the MMG, while the second row compares depth maps w/o and w/o the image modality.

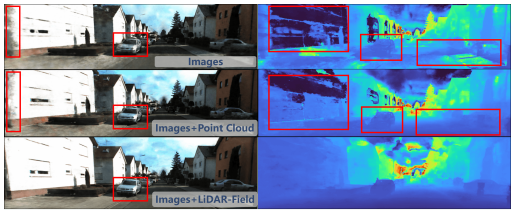

Figure 7: **Qualitative NVS results with GT-poses.** MUP outperforms single-modal methods i-NGP w/ and w/o point clouds and LiDAR-NeRF. Our method achieves significantly better depth estimation and NVS quality.

5.2 Comparison and Ablation in Pose-free Setting

Baselines. In pose-free settings, all baselines utilize multimodal inputs. Methods are divided into two distinct categories: one where registration is performed prior to reconstruction, and another where pose-free reconstruction simultaneously estimates the poses. Following prior pose-free NVS studies [16, 19, 48, 3, 10, 45], reconstruction is typically performed on short sequences without real-time constraints. Consequently, incremental SLAM systems are not directly comparable, and directly frame-to-frame registration is adopted instead [48].

For the first category, the registration method Colored-ICP [24] integrates both point clouds and images. We utilize it for registration and subsequently apply Alignmif [38] for multimodal reconstruction. Notably, For the second category, Nope-NeRF takes images input and leverages DPT [57] to construct pseudo point clouds. Moreover, we reformulate the Alignmif method in a pose-free framework, allowing gradient propagation to the poses for optimization. For all pose-free methods, we adopt the strategy from [45, 48] to obtain the poses of test views for rendering.

Comparison in NVS and Pose Estimation. The quantitative and qualitative results are presented in Table 1 and Fig. 5. Our method outperforms all approaches in both modalities. Previous pose-free methods Nope-NeRF [3] primarily targeted small-scale scenarios, and only can perform images NVS. Consequently, when applied to autonomous driving environments, Nope-NeRF underperforms due to the lack of scale and the imprecision in depth estimation. Alignmif [38] cannot be effectively used in ill-conditioned optimization. Its complex structure with multiple independent hash grids prevents efficient registration and reconstruction. As for the registration-first approach, Colored-ICP [24] exhibits limited accuracy in large-scale outdoor scenes. Our method achieves the highest pose estimation accuracy.

Ablation Study in pose-free setting. All ablation studies are conducted on KITTI-360 [15]. We firstly exclude the images and perform reconstruction using only the LiDAR point clouds. Table 2 underscores the critical role of multimodal fusion in enhancing accuracy. Besides, as shown in Fig. 6, the incorporation of images mitigates the limited LiDAR field of view, thereby enabling the acquisition

of depth maps (or point clouds) with a broader FoV. It also introduces complementary information beyond LiDAR's perspective, enhancing registration accuracy. We also conduct ablation studies on MMG module. The results indicate that relying solely on NeRF's implicit pose optimization fails to achieve accurate pose estimates and leads to convergence at local optima. Besides, we conduct ablation studies on image enhancement in MMG and the modality-specific C2F training strategy (MSC2F). All modules demonstrate effectiveness in pose-free experiments.

5.3 Ablation Study with Ground Truth Poses

In the pose-free setting, the precision of the estimated poses and the efficacy of our MMG module are pivotal to performance. Therefore, we exclude the MMG and conduct ablation experiments with GT pose to further demonstrate the advantages of our Unified NeRF with the multimodal-specific coarse-to-fine training strategy (MSC2F) and the consistency constraint in multimodal fusion. Additionally, to further demonstrate the effectiveness of our multimodal approach, We conduct comparative experiments with the single-modality LiDAR-NeRF [37] and i-NGP [21], where i-NGP is tested both with and without utilizing discrete LiDAR point clouds as depth supervision.

Ablation of MSC2F and Consistency Loss. Table 2 presents the ablation results on MSC2F and consistency constraint under the GT-Pose setting to verify the effectiveness of our method. By using a reprojection operation to link geometry and color, our method effectively ensures geometry-color coherence, resulting in improved reconstruction quality in both the image and point cloud NVS.

Comparision with Single-modality Methods. The quantitative and qualitative results are presented in Table 3 and Fig. 7. Our MSC2F fusion approach, along with the color-depth consistency constraint, effectively integrates features from both modalities. Thus, compared to single-modality methods and i-NGP [21] that with and without point clouds for depth supervision, we achieve high-quality NVS and the best results across both modalities.

6 Limitation

MUP demonstrates strong performance in pose-free multimodal NVS and pose estimation under challenging large-scale scenes. However, it is primarily designed for sensor data within a sequence and relies on temporal correlations between frames. Additionally, it is not designed to handle dynamic scenes, which is a non-negligible limitation in autonomous driving scenarios.

7 Conclusion

We revisit the limitations of single-modality pose-free methods in large-scale scenes. Subsequently, we propose a novel framework for Multimodal Unified Pose-free LiDAR-Camera NVS. Benefiting from the unified neural representation with MSC2F training strategy, the color-depth consistency constraint, the MMG module, and most importantly, the integration of different modalities and our pose optimization approach, we achieve geometry-aware, modality-consistent, pose-free reconstruction.

8 Acknowledgement

This work was supported by the National Key Research and Development Program of China (No. 2024YFE0211000), in part by the National Natural Science Foundation of China (No. 62372329, 62402341), in part by the Shanghai Scientific Innovation Foundation (No. 23DZ1203400), in part by the China Postdoctoral Science Foundation (No. BX20250383, GZB20250385, 2025M771530, 2025M771539, GZC20241225, 2025M771513), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[3] 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*, pages 4160–4169, 2023.

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022.

[7] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24–34, 2023.

[8] Junyuan Deng, Qi Wu, Xieyuanli Chen, Songpengcheng Xia, Zhen Sun, Guoqing Liu, Wenxian Yu, and Ling Pei. Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8218–8227, 2023.

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

[10] Hwan Heo, Taekyung Kim, Jiyoung Lee, Jaewon Lee, Soohyun Kim, Hyunwoo J Kim, and Jin-Hwa Kim. Robust camera pose refinement for multi-resolution hash encoding. In *International Conference on Machine Learning*, pages 13000–13016. PMLR, 2023.

[11] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19774–19783, 2023.

[12] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. *arXiv preprint arXiv:2305.01643*, 2023.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Yixing Lao, Xiaogang Xu, Xihui Liu, Hengshuang Zhao, et al. Corresnerf: Image correspondence priors for neural radiance fields. *Advances in Neural Information Processing Systems*, 36: 40504–40520, 2023.

[15] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

[16] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021.

[17] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9377–9384. IEEE, 2023.

[18] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023.

[19] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16539–16548, 2023.

[20] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

[22] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[23] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.

[24] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 143–152, 2017.

[25] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T Barron, and Ricardo Martin-Brualla. Camp: Camera preconditioning for neural radiance fields. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[27] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[28] François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat. Comparing icp variants on real-world data sets: Open-source library and experimental protocol. *Autonomous robots*, 34:133–148, 2013.

[29] Srikumar Ramalingam and Yuichi Taguchi. A theory of minimal 3d point to 3d plane registration and its generalization. *International journal of computer vision*, 102:73–90, 2013.

[30] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022.

[31] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022.

[32] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001.

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

[34] Liang Song, Guangming Wang, Jiuming Liu, Zhenyang Fu, Yanzi Miao, et al. Sc-nerf: Self-correcting neural radiance field with sparse views. *arXiv preprint arXiv:2309.05028*, 2023.

[35] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021.

[36] Zhen Tan, Zongtan Zhou, Yangbing Ge, Zi Wang, Xieyuanli Chen, and Dewen Hu. Td-nerf: Novel truncated depth prior for joint camera pose and neural radiance field optimization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 372–379. IEEE, 2024.

[37] Tang Tao, Longfei Gao, Guangrun Wang, Peng Chen, Dayang Hao, Xiaodan Liang, Mathieu Salzmann, and Kaicheng Yu. Lidar-nerf: Novel lidar view synthesis via neural radiance fields. *arXiv preprint arXiv:2304.10406*, 2023.

[38] Tang Tao, Guangrun Wang, Yixing Lao, Peng Chen, Jie Liu, Liang Lin, Kaicheng Yu, and Xiaodan Liang. Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21230–21240, 2024.

[39] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023.

[40] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12375–12385, 2023.

[41] Chen Wang, Jiadai Sun, Lina Liu, Chenming Wu, Zhelun Shen, Dayan Wu, Yuchao Dai, and Liangjun Zhang. Digging into depth priors for outdoor neural radiance fields. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1221–1230, 2023.

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[43] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[45] 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.

[46] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, pages 3–15. Springer, 2023.

[47] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023.

[48] Weiyi Xue, Zehan Zheng, Fan Lu, Haiyun Wei, Guang Chen, and Changjun Jiang. Geonlf: Geometry guided pose-free neural lidar fields. *arXiv preprint arXiv:2407.05597*, 2024.

[49] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023.

[50] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023.

[51] 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)*, pages 1323–1330. IEEE, 2021.

[52] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022.

[53] Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, and Xing Tang. Ray priors through reprojection: Improving neural radiance fields for novel view extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18376–18386, 2022.

[54] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. *arXiv preprint arXiv:2304.14811*, 2023.

[55] 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*, pages 586–595, 2018.

[56] Zehan Zheng, Fan Lu, Weiyi Xue, Guang Chen, and Changjun Jiang. Lidar4d: Dynamic neural fields for novel space-time view lidar synthesis. *arXiv preprint arXiv:2404.02742*, 2024.

[57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper does discuss the limitations of the work performed by the authors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: This paper provide open access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: This paper specify all the training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This paper report appropriate information about the statistical significance of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conducted in the paper conform, in evrery respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While essential for autonomous driving, reconstruction and closed-loop simulation raise privacy and employment concerns, highlighting the need to balance technological progress with societal responsibility.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mentioned creators or original owners of assets and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper introduces new assets well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not involve research related to LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.