AlignMiF: Geometry-Aligned Multimodal Implicit Field for LiDAR-Camera Joint Synthesis

Tang Tao[1] Guangrun Wang[3] Yixing Lao[4] Peng Chen[5] Jie Liu[6] Liang Lin[7] Kaicheng Yu[8] †
Xiaodan Liang[1,2] †
[1] Shenzhen Campus of Sun Yat-sen University [2] DarkMatter AI Research [3] University of Oxford
[4] HKU [5] Cainiao Group [6] NCUT [7] Sun Yat-sen University [8] Westlake University
{trent.tangtao, kaicheng.yu.yt, hxdliang328}@gmail.com

Abstract

Neural implicit fields have been a de facto standard in novel view synthesis. Recently, there exist some methods exploring fusing multiple modalities within a single field, aiming to share implicit features from different modalities to enhance reconstruction performance. However, these modalities often exhibit misaligned behaviors: optimizing for one modality, such as LiDAR, can adversely affect another, like camera performance, and vice versa. In this work, we conduct comprehensive analyses on the multimodal implicit field of LiDAR-camera joint synthesis, revealing the underlying issue lies in the misalignment of different sensors. Furthermore, we introduce AlignMiF, a geometrically aligned multimodal implicit field with two proposed modules: Geometry-Aware Alignment (GAA) and Shared Geometry Initialization (SGI). These modules effectively align the coarse geometry across different modalities, significantly enhancing the fusion process between LiDAR and camera data. Through extensive experiments across various datasets and scenes, we demonstrate the effectiveness of our approach in facilitating better interaction between LiDAR and camera modalities within a unified neural field. Specifically, our proposed AlignMiF, achieves remarkable improvement over recent implicit fusion methods (+2.01 and +3.11 image PSNR on the KITTI-360 and Waymo datasets) and consistently surpasses single modality performance (13.8% and 14.2% reduction in LiDAR Chamfer Distance on the respective datasets). Code release: https://github.com/tangtaogo/alignmif.

1. Introduction

Synthesizing novel views has recently seen significant progress due to Neural Radiance Field (NeRF) [26], which models a 3D scene as a continuous function and leverages differentiable rendering, resulting in a de facto standard to render novel views. Notably, the recent NeRF methods have shown impressive performance on downstream tasks such as autonomous diving [43, 50, 52, 53]. In such practical scenarios, both images and LiDAR sensors are typically utilized. Currently, researchers extend the NeRF formulation for novel LiDAR view synthesis [15, 42, 60], which treat the oriented LiDAR laser beams as a set of rays and render 3D points and intensities in a similar fashion as RGB.

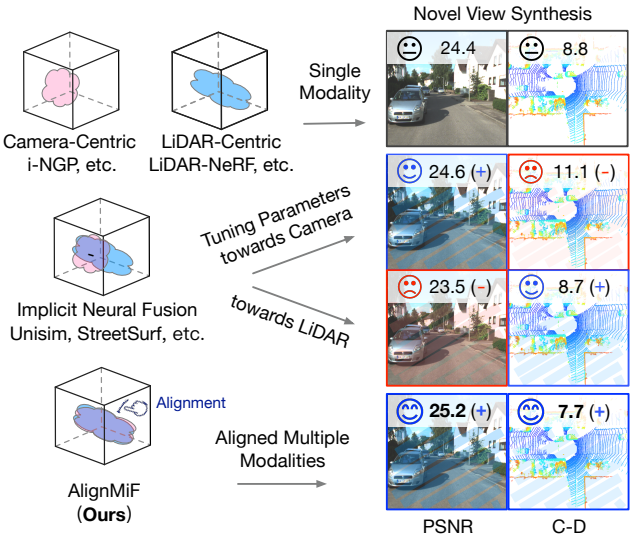

Figure 1. **The misalignment issue in multimodal implicit field.** For implicit neural fusion, there is a trade-off between the modalities due to the misalignment, making it challenging to improve both modalities simultaneously. Conversely, our method addresses the misalignment issue and achieves boosted multimodal performance. The metrics are PSNR and Chamfer Distance (C-D).

However, the exploration of multimodal learning in NeRF is still in its early stages. There are initial attempts, such as UniSim [55] and NeRF-LiDAR [60], to incorporate multimodal inputs through implicit fusion, i.e., sharing the implicit features in one single field, aiming to lever-

†Co-corresponding author.

age the complementary information from different modalities to enhance NeRF's capabilities. Accordingly, the integration of multiple input modalities is expected to boost model performance, but our results show that the naive multimodal NeRF, which relies on direct implicit fusion, does not outperform its unimodal counterpart. As illustrated in Fig. 1, it is challenging to improve both modalities simultaneously. Intuitively, for NeRF optimization, incorporating more information is expected to lead to better results [7, 18, 32, 44, 59], which is not fully realized in current multimodal fields when fusing LiDAR and camera modalities. These observations highlight the need for ongoing research and advancements in multimodal learning in NeRF.

In this study, we perform comprehensive analyses of multimodal NeRF that integrates LiDAR and camera sensors for joint synthesis. Our preliminary experiments, conducted on real-world datasets such as KITTI-360 [23] and Waymo [38], reveal that different modalities often contradict each other. However, such conflicts are not observed in the synthetic AIODrive dataset [49]. These findings lead us to speculate that the underlying issue lies in the misalignment of different modalities, e.g., spatial misalignment and temporal misalignment. When modalities are not properly aligned, the implicit fusion of conflicting information can hinder network optimization, resulting in suboptimal outcomes for both modalities. To further investigate and validate this misalignment issue, we conducted extensive analyses, including the examination and visualization of raw sensor inputs, hash grid features, and the density values from the geometry network. Moreover, we conducted experimental analyses on various network architecture designs to validate our findings. These investigations provide valuable insights into the misalignment issue and its implications on the performance of the unified multimodal NeRF.

To tackle the challenge of misalignment in multimodal NeRF, we propose a twofold solution, called AlignMiF, to align the consistent coarse geometry across different modalities while keeping their individual detail characteristics. Firstly, we decompose the hash encoding to allow each modality to concentrate on its own information. We then apply an alignment constraint at the coarse geometry levels, facilitating mutual enhancement and cooperation between modalities, referred to as the Geometry-Aware Alignment (GAA) module. Secondly, we utilize the hash grid features from a pre-trained field as a share initialization of the geometry, referred to as the Shared Geometry Initialization (SGI) module. This shared initialization further enhances the alignment process, allowing each modality to capture its respective details upon it. Both GAA and SGI aim to align the coarse geometry while preserving their unique details.

Through comprehensive experiments conducted on multiple datasets and scenes, we validate the effectiveness of our approach in boosting the interaction between LiDAR and camera modalities within a unified framework. Our proposed modules, GAA and SGI, contribute to improved alignment and fusion, leading to enhanced performance and more accurate synthesis of novel views. Specifically, as a result, our AlignMiF achieves remarkable improvement over the implicit fusion (e.g., +2.01 and +3.11 PSNR on KITTI-360 and Waymo datasets) and consistently outperforms the single modality (e.g., 13.8% and 14.2% reduction in LiDAR Chamfer Distance on the respective datasets).

Overall, our contributions are as follows:

- We perform comprehensive analyses of multimodal learning in NeRF, identifying the modality misalignment issue.
- We propose AlignMiF, with GAA and SGI modules, to address the misalignment issue by aligning the consistent coarse geometry of different modalities while preserving their unique details.
- We demonstrate the effectiveness of our method quantitatively and qualitatively through extensive experiments conducted on multiple datasets and scenes.

2. Related Work

2.1. NeRF for Novel View Synthesis

Neural Radiance Fields (NeRF) [26] have revolutionized the long-standing novel view synthesis. Many NeRF variants have been proposed, focusing on aspects such as acceleration [5, 27, 56], anti-aliasing [2, 3, 13], managing casual camera trajectories [25, 45], and generalization capabilities [16, 57]. There also emerges research leveraging depth information for view synthesis [7, 28, 32, 44]. In parallel, great progress has been made in NeRF applications for handling complex and large-scale environments such as urban outdoor scenes [20, 24, 31, 40, 43, 50, 52–54]. Concurrently, researchers extend the NeRF formulation for novel LiDAR view synthesis [14, 15, 42, 60, 63], which treat the oriented LiDAR laser beams in a similar manner to camera rays. Given the recent advancements for novel view synthesis in different modalities, in this work, we dig into the investigation of the unified multimodal NeRF framework.

2.2. Multimodal Learning in NeRF

Recent works have also explored multi-task learning in NeRF, which involves synthesizing panoptic or semantic labels alongside RGB views [10, 62, 64]. However, the exploration of multimodal learning in NeRF is still in its early stages. Some current works have attempted to incorporate multimodal inputs, such as NeRF-LiDAR [60], which leverages images, semantic labels, and LiDAR data to generate LiDAR points and corresponding labels. Another work, StreetSurf [11], utilizes LiDAR as supervision for street view multi-view reconstruction and can also generate 3D points. UniSim [55], on the other hand, is a neural sensor simulator that takes multi-sensor inputs into a shared

implicit field, and simulates LiDAR and camera data at new viewpoints. These preliminary efforts primarily focus on simple implicit fusion, i.e., directly sharing the implicit features of different modalities in a single field. However, we actually find that modalities in this multimodal NeRF often contradict each other and cannot outperform its unimodal counterpart. Our work delves deeper into exploring the intricate interactions between multimodalities and proposes a geometry-aligned multimodal implicit field.

3. Problem Analysis

3.1. Preliminaries

NeRF [26] models a scene as a continuous volumetric field. Given a 3D location $\mathbf{x}$ and a viewing direction $\boldsymbol{\theta}$ as input, NeRF learns an implicit function f that predicts the volume density σ and color $\mathbf{c}$ as $(\sigma, \mathbf{c}) = f(\mathbf{x}, \boldsymbol{\theta})$. Specifically, given rays $\mathbf{r}$ originated from camera origin $\mathbf{o}$ in direction $\mathbf{d}$, i.e., $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, the corresponding pixel color is approximated by the numerical quadrature of the color $\mathbf{c}_i$ and density σ_i of samples along the ray: $\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i\big(1 - \exp(-\sigma_i\delta_i)\big)\mathbf{c}_i$, where $T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)$ and δ_i is the distance between adjacent samples.

More recent works [15, 42, 60] extend the traditional NeRF to LiDAR sensor, treating the oriented LiDAR laser beams as a set of rays. Slightly abusing the notation, let $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ be a ray casted from the LiDAR sensor, where $\mathbf{o}$ denotes the LiDAR center, and $\mathbf{d}$ represents the normalized direction vector of the corresponding beam. Then the depth measurement $\hat{D}(\mathbf{r})$ can be approximated by calculating the expectation of the samples along the ray: $\hat{D}(\mathbf{r}) = \sum_{i=1}^{N} T_i\big(1 - \exp(-\sigma_i\delta_i)\big)t_i$. The view-dependent features of LiDAR, including the intensities and ray-drop probabilities, can be rendered similarly to RGB color.

3.2. Multimodal Learning in NeRF

Building on the NeRF formulations for different sensors mentioned above, there have been preliminary works, such as UniSim [55] and NeRF-LiDAR [60], integrating these into a unified multimodal NeRF framework. These methods directly share implicit features across different modalities, and the optimization targets can be combined as:

$$\mathcal{L}_{\text{total}} = \lambda_l \mathcal{L}_{\text{LiDAR}}(\mathbf{r_l}) + \lambda_c \mathcal{L}_{\text{camera}}(\mathbf{r_c}), \qquad (1)$$

where $\mathbf{r_l} \in R_l$ and $\mathbf{r_c} \in R_c$ are the sensor training rays, and λ are weight coefficients to balance each term. However, current efforts only primarily focus on simple implicit fusion, and the multimodal NeRF has not been fully exploited.

3.3. The Misalignment Issue

Multimodal learning helps to comprehensively understand the world, by integrating different senses. Accordingly, multiple input modalities are expected to boost model performance. However, our findings suggest that the current multimodal NeRF cannot surpass its uni-modal counterpart in terms of performance. As shown in Tab. 1, regardless of how we adjust the weights of the two modalities, we find that it is challenging to improve both modalities simultaneously (more tuning results are shown in Fig. 10 of the appendix). Indeed, similar challenges have also been observed in other multimodal research areas. Previous researchers claimed that different modalities exhibit inconsistent representations and tend to converge at different rates, leading to uncoordinated convergence problems [30, 39, 46]. However, these theories are not applicable to our scenario, as our representation is a unified field that reflects the real world. Intuitively, for NeRF optimization, it is expected that having more information would lead to better results [7, 18, 28, 32, 59]. We further conducted experiments on the synthetic dataset, AIODrive [49], which was collected from CARLA Simulator [8]. As shown in the bottom block of Tab. 1, we can observe that mutual boosting between the two modalities could be achieved on the perfect synthetic data. Based on these observations, we speculate that the underlying issue lies in the misalignment of modalities. When two modalities are not properly aligned, the im-

Table 1. **The misalignment issue.** When directly implicit fusing the multiple modalities cannot outperform their unimodal counterparts simultaneously in real-world datasets. While this challenge is not present in the synthetic dataset. Here, w denotes the weight ratio between the two modalities, $w_\lambda = \lambda_c/\lambda_l$.

Method	w_λ	RGB Metric		LiDAR Metric	
		PSNR↑	SSIM↑	C-D↓	F-score↑
KITTI-360 *(real-world)*					
i-NGP [27]	–	24.45	0.787	–	–
LiDAR-NeRF [42]	–	–	–	0.088	0.920
UniSim-SF [55]	0.1	23.54 (–)	0.759	0.087 (+)	0.929
	0.5	24.38 (–)	0.792	0.100 (–)	0.920
	2.0	24.80 (+)	0.809	0.124 (–)	0.900
Waymo *(real-world)*					
i-NGP [27]	–	28.20	0.830	–	–
LiDAR-NeRF [42]	–	–	–	0.179	0.885
UniSim-SF [55]	0.5	26.41 (–)	0.789	0.172 (+)	0.891
	1.0	27.12 (–)	0.805	0.181 (–)	0.885
	5.0	28.33 (+)	0.830	0.227 (–)	0.840
AIODrive *(synthetic)*					
i-NGP [27]	–	34.43	0.893	–	–
LiDAR-NeRF [42]	–	–	–	0.178	0.873
UniSim-SF [55]	0.1	34.25 (–)	0.901	0.138 (+)	0.921
	1.0	34.53 (+)	0.904	0.153 (+)	0.915
	5.0	34.64 (+)	0.905	0.191 (–)	0.914

plicit fusion of conflicting information causes the network to struggle to determine the correct optimization direction, resulting in suboptimal results for both.

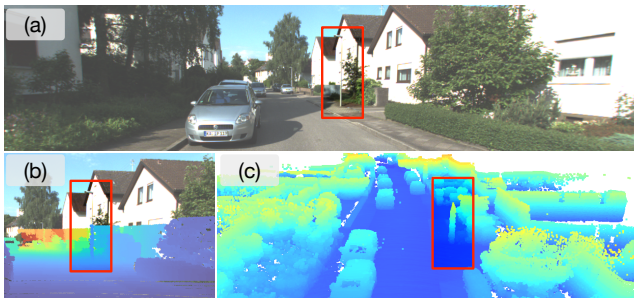

Figure 2. **Analysis of misalignment from raw sensor inputs.** (a) Original image, (b) Image with projected points from associate LiDAR frame, (c) LiDAR points of all scene frames. As highlighted in the red box, the observations obtained from LiDAR and camera sensors for the same pole are distinct (zoom-in for better views).

To further investigate the misalignment issue, we begin with the raw modality inputs. As illustrated in Fig. 2, both the camera and LiDAR sensors essentially represent the same overall scene, yet they exhibit variances in capturing finer details. For example, when scanning the same lamp post, the pole obtained from LiDAR appears to have larger diameter compared to the one captured by the RGB camera. In fact, the perceptual characteristics of these sensors are inherently different. LiDAR lacks semantic perception of objects and provides rougher boundaries, while the camera lacks distance perception. Moreover, even without considering calibration errors between multiple sensors, inherent systematic errors exist in each sensor [48, 58], e.g., the different operating frequency and trigger mechanism of camera, LiDAR, GPS, and IMU. Consequently, it becomes challenging to align all details across the two modalities, resulting in ambiguous conflicts within one unified field.

Next, we investigate the learned hash gird features of different modalities. The multi-resolution hash encoding introduced by iNGP [27] is expressive and efficient, which is a common practice in current works [11, 42, 55]. In Fig. 3, we visualize the learned hash features on the x-y plane. It is apparent that the learned geometry from LiDAR is well represented, and the shape of the hash features aligns with the scene. Then, both modalities primarily focus on their respective field of view (FOV), with the camera capturing the top-left part due to its front-facing orientation. Notably, even in the overlap area of the FOV, the highlighted feature regions of interest differ, indicating conflicting ambiguity between the two modalities. Consequently, when utilizing the simple implicit fusion, i.e., directly sharing the implicit features of different modalities, the model becomes confused by the misaligned modalities, resulting in disorganized hash features (bottom row in Fig. 3), ultimately yielding suboptimal results for both.

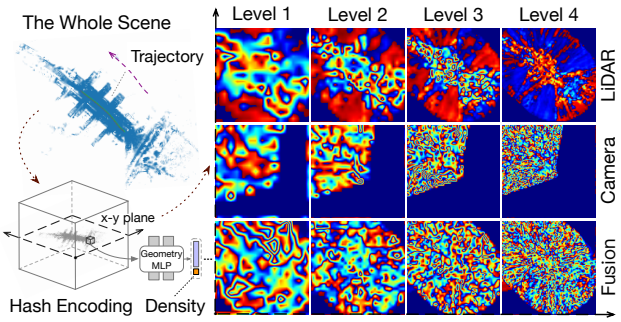

Figure 3. **Analysis of misalignment from bird's eye view hash grid features.** We show the first 4 levels of the hash features on the x-y plane. The camera is front-facing along the trajectory and brighter or more saturated colors represent higher feature values.

Furthermore, we present additional analysis and experiments in Fig. 5 and Tab. 4. These findings both provide valuable evidence to support our claims and contribute to a deeper understanding of the misalignment issue.

4. AlignMiF

In this section, we introduce our multimodal implicit field, AlignMiF, with two proposed geometrical alignment modules, aiming to mitigate the misalignment problem and enhance performance across both modalities.

4.1. Geometry-Aware Alignment

To alleviate the misalignment issue between the two modalities, we first decompose the hash encoding, allowing each modality to focus on its own information; while subsequently, we need to enhance information interactions between modalities. Inspired by Neuralangelo [21] and HR-Neus [22], we acknowledge that different levels of hash encoding correspond to different levels of fine-grained geometry information. Specifically, the lower-indexed grid levels contain the most information about the coarse geometry, while the higher-level grids primarily contain information about high-frequency details. As observed in our earlier analysis, the misalignment primarily occurs at the detailed levels, while both modalities share the same underlying coarse scene geometry of the real world. Building upon this observation, we further propose a **G**eometry-**A**ware **A**lignment (GAA), which specifically aligns the two modalities at the coarse geometry levels, facilitating mutual enhancement and cooperation between them.

Specifically, we denote the multi-resolution hash encoding with total L levels and a feature dimension of d as γ. Given a query point $\mathbf{x}$, the 3D feature grid at each level is first trilinearly interpolated and then concatenated together to form the final feature vector: $\gamma(\mathbf{x}) = \overline{F_1 F_2 \dots F_L} \in \mathbb{R}^{L \times d}$, where $F_l \in \mathbb{R}^d$ represents the interpolated feature at level l. Then as previous works [21, 22], we expand the definition of γ to take in an additional parameter β that helps

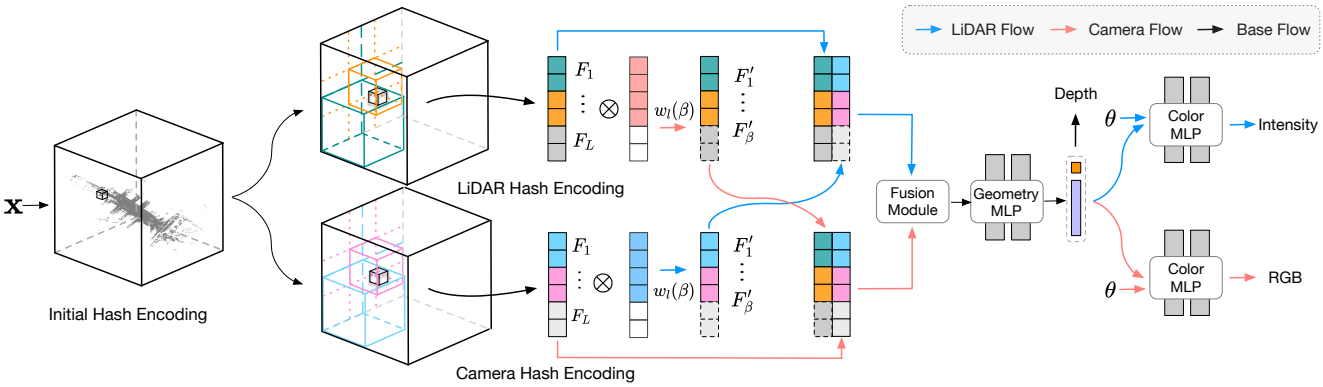

Figure 4. **The illustration of our AlignMiF framework.** The proposed Geometry-Aware Alignment (GAA) of the decomposed hash encoding and the Shared Geometry Initialization (SGI) are incorporated together to tackle the misalignment issue.

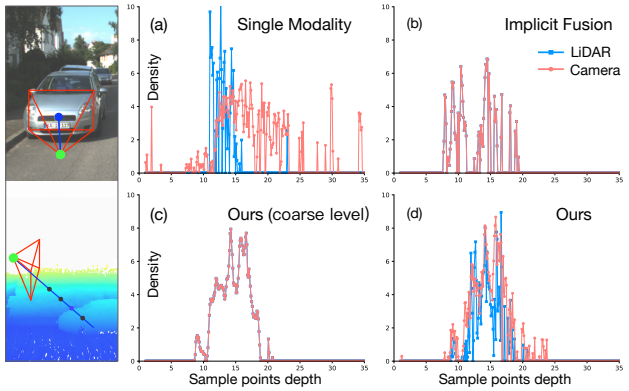

Figure 5. **Analysis of misalignment from the density values and qualitative analysis of our propose GAA.**

zero out the higher-level resolution feature grid:

$$\gamma(\mathbf{x},\beta) = \overline{F'_1 F'_2 \dots F'_L}, \text{ where } F'_l = w_l(\beta) F_l,$$
$$w_l(\beta) = \frac{1-\cos(\pi \cdot \text{clamp}(\beta - l + 1, 0, 1))}{2} \quad (2)$$

Intuitively, grid layer F_l will be fully activated if $l \leq \beta$ and all other higher grid layers will be zeroed out. Note that all grid layers are available when β is not provided. We denote the two decomposed hash encoding as γ_{lidar} and γ_{camera}, and then our GAA can be formulated as:

$$\psi_{GAA}(\mathbf{x},\beta) = \begin{cases} \mathcal{F}(\gamma_{lidar}(\mathbf{x}), \gamma_{camera}(\mathbf{x},\beta)) & \mathbf{x} \sim \mathbf{r_l} \\ \mathcal{F}(\gamma_{lidar}(\mathbf{x},\beta), \gamma_{camera}(\mathbf{x})) & \mathbf{x} \sim \mathbf{r_c} \end{cases}, \quad (3)$$

where $\mathcal{F}$ denotes the fusion module for the alignment, e.g., concatenation or attention mechanism. Shortly, our GAA combines the hash features of the current modality with the aligned coarse level hash features from another modality and the illustration is on Fig. 4.

In Fig. 5, we visualize the learned density values from the geometry MLP for sampled points along the ray. In (a), the density values, reflecting the geometry, learned by the LiDAR and camera are very inconsistent, which correspond to the previously obtained hash features in Fig. 3. In (b), when employing the implicit fusion, the learned densities can be confounded by misalignment between the two modalities, resulting in erroneous geometry. Conversely, in (c) and (d), our method achieves a more robust and comprehensive geometry representation that incorporates information from both modalities. Specifically, the GAA module aligns the coarse level of geometry, i.e., the consistent densities, as shown in (c), while both modalities capture their respective finer details as in (d), which demonstrates the effectiveness of our method.

4.2. Shared Geometry Initialization

In the visualizations of the hash features and density values shown in Fig. 3 and Fig. 5, it is evident that the learned geometry from the camera can be inaccurate. This observation also conformed with the shape-radiance ambiguity discussed in previous works [7, 28, 44]. Although we propose a geometry-aware alignment module to enhance the camera's geometry using LiDAR information, the learning and alignment process still remain implicit. To address this, we propose utilizing the hash grids from a pre-trained field with rough geometry, e.g., the trained LiDAR field, as a shared geometry initialization for both modalities. Previous works [4, 55] have also suggested using LiDAR to constrain the volume grids. However, considering the different FOV and the varying details of LiDAR and camera data as analyzed earlier, we additionally learn the hash features of both modalities after the initialized hash encoder, rather than relying only on LiDAR. Then we directly add hash features from the initial encoding to each modality as the shared geometry information. Specifically, we denote the shared initialized hash grid as γ_{init}, and the proposed **S**hared **G**eometry **I**nitialization (SGI) can be formulated as:

$$\phi_{SGI}(\mathbf{x}) = \begin{cases} \gamma_{init}(\mathbf{x}) + \gamma_{lidar}(\mathbf{x}) & \mathbf{x} \sim \mathbf{r_l} \\ \gamma_{init}(\mathbf{x}) + \gamma_{camera}(\mathbf{x}) & \mathbf{x} \sim \mathbf{r_c} \end{cases}. \quad (4)$$

As shown in Fig. 6, after applying our SGI module, we observe remarkable improvements in the hash features of both LiDAR and camera encoding, compared with Fig. 3. Especially, the camera's hash features exactly focus on the relevant regions of the scene geometry. These observations demonstrate the effectiveness of our proposed module.

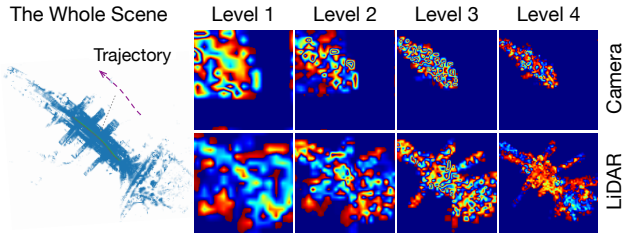

Figure 6. **Qualitative analysis of our propose SGI module.**

In summary, our SGI leverages the initialization from a pre-trained field as the shared coarse geometry while still allowing the learning of hash features from both modalities to capture their respective details. This aligns with the GAA module, and both of them align coarse geometry while preserving modalities' unique characteristics.

4.3. AlignMiF Formulation

Overall, as illustrated in Fig. 4, combining the proposed two simple yet effective modules, we summarize the formulation of our AlignMiF as follows:

$$\begin{aligned}\Psi_{GAA}^{SGI}(\mathbf{x},\beta) &= \psi_{GAA}(\mathbf{x},\beta;\ \phi_{SGI}) \\ &= \begin{cases}\mathcal{F}(\phi_{lidar}(\mathbf{x})),\ \phi_{camera}(\mathbf{x},\beta)) & \mathbf{x}\sim\mathbf{r_l} \\ \mathcal{F}(\phi_{lidar}(\mathbf{x},\beta),\ \phi_{camera}(\mathbf{x})) & \mathbf{x}\sim\mathbf{r_c}\end{cases},\end{aligned} \tag{5}$$

where $\phi_{SGI}(\mathbf{x},\beta) = \gamma_{init}(\mathbf{x}) + \gamma(\mathbf{x},\beta)$.

5. Experiment

5.1. Experimental Setting

Datasets. We conducted experiments on three datasets: one synthetic dataset, AIODrive [49], and two challenging real-world datasets, KITTI-360 [23] and Waymo Open Dataset [38]. These datasets were collected using RGB cameras and LiDAR sensors.

Evaluation metrics. For novel image view synthesis, following the previous methods [26, 27], our evaluations are based on three widely-used metrics, *i.e.*, peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [47], and the learned perceptual image patch similarity (LPIPS) [61]. For novel LiDAR view synthesis, as works [15, 42], we report the Chamfer Distance (C-D) between the rendered and original LiDAR point clouds and the F-Score with a threshold of 5cm. The novel intensity image is evaluated using mean absolute error (MAE).

Baselines. For the uni-modal model, we consider the popular i-NGP [27] for novel image view synthesis and the concurrent LiDAR-NeRF [42] for novel LiDAR view synthesis. For multimodal evaluation, we use UniSim [55] as the main baseline. Since this work focuses on investigating the relationship between multimodalities, we specifically re-implement its implicit fusion component and also did not consider dynamic foreground. To distinguish it from the original UniSim, we refer to it as UniSim-Static-Implicit Fusion, abbreviated as UniSim-SF. Details of dataset sequences and splits and implementation details are provided in supplementary materials.

5.2. Main Results

Results on KITTI-360 and Waymo dataset. In Tab. 2, we present the evaluation results on the KITTI-360 and Waymo datasets. As mentioned in Sec. 3.3, the UniSim-SF model fails to achieve simultaneous improvements in multimodal performance, and there is a trade-off between the modalities. To ensure a fair comparison, we carefully fine-tuned the parameters to ensure that the LiDAR or camera modality in UniSim-SF surpassed its corresponding single-modality counterpart by a small margin, preventing either modality from being significantly worse. Compared with the carefully tuned UniSim-SF model and the corresponding single-modality models, our AlignMiF achieves superior results by clear margins. Specifically, our method achieves remarkable improvement over UniSim-SF by +2.01 and +3.11 image PSNR on KITTI-360 and Waymo datasets respectively, and outperforms the single-modality models overall, as evidenced by the 13.8% and 14.2% reduction in LiDAR Chamfer Distance respectively. With two proposed alignment modules, our method facilitates better fusion of different modalities, resulting in more accurate understanding of the scene and improved results.

Furthermore, we provide qualitative results in Fig. 7, which clearly demonstrate the mutual benefits of our AlignMiF. As highlighted with the boxes, the LiDAR modality significantly enhances the learning of image and depth quality in the camera, while the semantic information from RGB aids the LiDAR in better converging to object boundaries. Please refer to our supplementary materials for all scene results and more visualization results.

Results compared with StreetSurf. In Tab. 3, we specifically compare the PSNR metric with StreetSurf [11] as it does not learn LiDAR intensity and ray-drop. It is worth noting that StreetSurf produces lower image results with the implicit LiDAR-camera fusion compared to the single-camera modality. This further emphasizes the misalignment issue across different representations and methods. In contrast, our method successfully improves the performance of both modalities and achieves state-of-the-art results.

Table 2. **Novel view synthesis on KITTI-360 dataset and Waymo dataset**. AlignMiF outperforms the baselines in all metrics.

Method	M	KITTI-360 Dataset						Waymo Dataset					
		RGB Metric			LiDAR Metric			RGB Metric			LiDAR Metric		
		PSNR↑	SSIM↑	LPIPS↓	C-D↓	F-score↑	MAE↓	PSNR↑	SSIM↑	LPIPS↓	C-D↓	F-score↑	MAE↓
i-NGP [27]	C	24.61	0.808	0.181	–	–	–	28.82	0.831	0.380	–	–	–
LiDAR-NeRF [42]	L	–	–	–	0.094	0.916	0.122	–	–	–	0.197	0.871	0.040
UniSim-SF [55]$^{\triangle}$	LC	23.30 (–)	0.758	0.268	0.090 (+)	0.924	0.097	26.67 (–)	0.788	0.417	0.186 (+)	0.878	0.039
UniSim-SF [55]$^{\triangledown}$	LC	24.94 (+)	0.812	0.184	0.114 (–)	0.906	0.095	28.98 (+)	0.833	0.374	0.355 (–)	0.786	0.045
AlignMiF (ours)	LC	25.31 (+)	0.826	0.164	0.081 (+)	0.928	0.099	29.78 (+)	0.845	0.339	0.169 (+)	0.885	0.038

M, L, C denotes modality, LiDAR, camera respectively. $\triangle$ and $\triangledown$ represent tuning parameters towards LiDAR and camera modality respectively.

Table 3. **Comparison with StreetSurf on Waymo dataset.**

Sequence	StreetSurf [11]		AlignMiF	
	C	LC	C	LC
seg1137922...	28.33	27.64	29.26	30.16
seg1067626...	29.01	27.68	29.52	30.27
seg1776195...	26.71	25.35	28.20	29.22
seg1172406...	28.50	27.86	28.30	29.47
Average	28.14	27.13 (–)	28.82	29.78 (+)

L, C denotes LiDAR, camera respectively and the metric is PSNR↑.

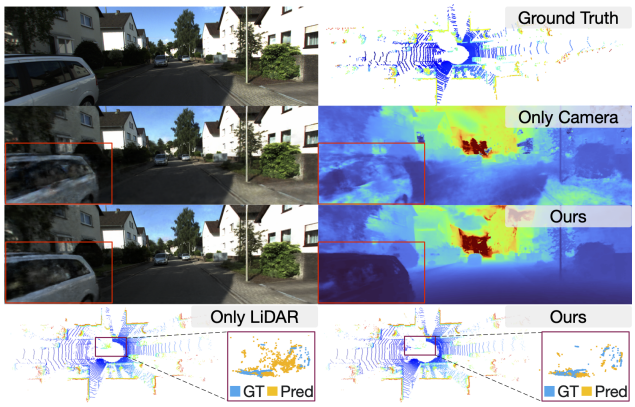

Figure 7. **Qualitative results on KITTI-360 dataset.** Our method achieves mutual boosting between both modalities (zoom-in for the best of views).

Table 4. **Analysis of misalignment from the network design.**

Method	RGB Metric		LiDAR Metric	
	PSNR↑	SSIM↑	C-D↓	F-score↑
Single Modality	24.45	0.787	0.088	0.920
Decompose Geometry-net	24.43	0.784	0.084	0.929
Decompose Densities	24.56	0.788	0.084	0.929
+ Hard Constraint	24.85	0.806	0.089	0.926
Decompose Hash-encoder	24.42	0.793	0.087	0.931
+ Hard Constraint	24.53	0.793	0.084	0.929

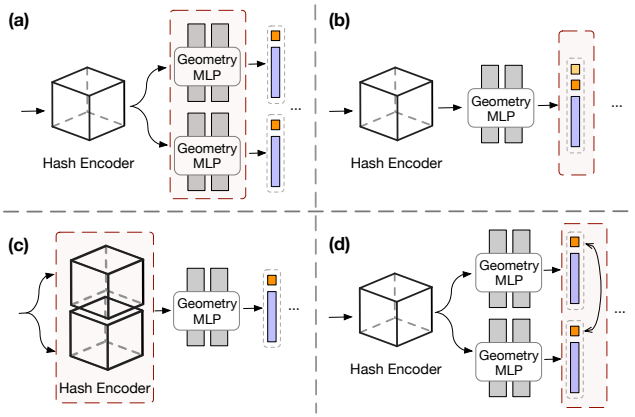

Figure 8. **Different designs of the network architecture.** (a) Decompose geometry net, (b) Decompose densities, (c) Decompose hash encoder, (d) Decompose geometry net with hard constraint.

5.3. Ablation Study

Analysis of the misalignment from the network design. During the early stages of exploring the multimodal problem, we conducted various experiments on the network architecture. We attempted to enhance the network's capacity by increasing the hash encoding resolution and feature dimensions, as well as adjusting the depth and width of the MLP network and increasing training iterations to handle multimodal inputs. However, these attempts did not yield significant improvements, as the underlying issue was not identified. Subsequently, we delved into the network structure design, as illustrated in Fig. 8 and the results are present on Tab. 4. In Fig. 8 (a), we decomposed the geometry MLP and surprisingly found that separating the geometry MLP resulted in almost no interference between the two modalities. However, this separation also meant that they did not mutually enhance each other. To further investigate the issue with the geometry net, in Fig. 8 (b), we designed a shared geometry MLP but allowed the network to separately output two density values for each modality. We observed that this approach also enabled independent learning for both modalities, leading us to identify the issue as the misalignment in the density values, or more specifically, geometry misalignment. Having identified the problem as the geometry misalignment, we proceeded to decompose the source hash decoding as Fig. 8 (c), and once again observed that the two modalities did not affect each other. Further-

Table 5. **Ablation study for the proposed components.**

Method	RGB Metric		LiDAR Metric	
	PSNR↑	SSIM↑	C-D↓	F-score↑
Single Modality	24.45	0.787	0.088	0.920
SGI				
Load LiDAR Encoder	24.34	0.784	0.096	0.920
+ Fix LiDAR Encoder	18.60	0.537	0.120	0.904
Detach RGB Density	12.41	0.597	0.098	0.916
GAA				
Addition	24.47	0.808	0.081	0.929
Attention	24.48	0.791	0.089	0.929
Concatenation	24.64	0.810	0.079	0.930
Share Coarse-Geo	24.45	0.794	0.087	0.930
AlignMiF				
w/ SGI	24.70 (+)	0.806	0.079 (+)	0.930
w/ GAA	24.64 (+)	0.810	0.079 (+)	0.930
AlignMiF	**25.20 (+)**	**0.816**	**0.077 (+)**	**0.932**

more, building upon the (b) and (c), we also tried to incorporate an alignment technique similar to the NeRF distillation work [9]. Specifically, we directly aligned the density values of the two separate networks as Fig. 8 (d). Although this hard constraint demonstrated some improvement in the outcomes, it remained a forced alignment and failed to address the underlying conflict. Consequently, the performance improvement was limited as in Tab. 4. Through these efforts in model network structure design, we not only verified the geometry misalignment issue but also inspired the design of our AlignMiF, which involves decomposing hash encoding and aligning coarse geometry.

Ablations on the proposed modules of AlignMiF. In Tab. 5, we conduct an ablation study to analyze the effects of the proposed components. Specifically, we observe that both the GAA and SGI modules improve the multimodal performance, which aligns with the qualitative analysis presented earlier in Fig. 5 and Fig. 6. Moreover, combining the two modules leads to further improvements, indicating that they effectively alleviate the misalignment issue.

We also provide an investigation of the design of these modules. For SGI, we explored different initialization strategies. One straightforward approach was to directly load a pre-trained LiDAR encoder as the initialization for geometry, referred to as *Load LiDAR Encoder*. However, this approach did not yield better results due to misalignment issues. Then we considered fixing the pre-trained geometry, i.e., *Fix LiDAR Encoder*, to mitigate the interference caused by noisy camera geometry, and only the MLPs were trained, similar to the training process in CLONeR [4]. Nevertheless, we got unfavorable results as the two modalities had different FOVs, resulting in the training being effective only in the pre-trained LiDAR FOV as shown in Fig. 12 of the appendix. Moreover, the analyzed misalignment depicted in Fig. 2 also contributes to obstacles. Additionally, we also attempted to detach the gradient of density from the camera modality to avoid geometry conflicts, which is similar to gradients blocking in Panoptic-Lifting [36], but the FOV mismatch and misalignment issue persisted. Hence, none of these designs proved as effective as our proposed SGI, which provides shared initial coarse geometry while allowing the learning of hash features from both modalities to capture their respective details.

For GAA, we explored different fusion strategies for alignment, including addition, concatenation, and attention mechanisms. We adopted the efficient attention structure from ER-NeRF [19]. Improvements can be observed with each fusion approach, and we ultimately selected concatenation as it's the most effective. Moreover, further exploration of more powerful fusion modules for alignment remains a promising research direction. Additionally, it's also considered to share the coarse geometry in GAA rather than aligning, however, the results obtained are not satisfactory. As also observed by Panoptic-Lifting [36], despite the underlying scene geometry being the same, features required for different modalities and representations might be slightly different, e.g., the LiDAR intensity and image color. Thus, similar to other works [33, 37], we employ a separate grid encoder for each specific modality.

6. Conclusion

In this paper, we thoroughly investigated and validated the misalignment issue in multimodal NeRF through various analyses, such as the examination and visualization of raw sensor inputs, hash features, and density values, as well as experiments on various network architectures. Futhermore, we propose AlignMiF, with two simple yet effective modules, Geometry-Aware Alignment (GAA) and Shared Geometry Initialization (SGI), to address the misalignment issue by aligning the consistent coarse geometry of different modalities while preserving their unique details. We conduct extensive experiments on multiple datasets and scenes and demonstrate the effectiveness of our proposed method in improving multimodal fusion and alignment within a unified NeRF framework. We hope that our work can inspire future research in the field of multimodal NeRF.

Acknowledgements. This work was supported in part by the National Key R&D Program of China under Grant No. 2020AAA0109700, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), Mobility Grant Award under Grant No. M-0461, Shenzhen Science and Technology Program (Grant No. RCYX20200714114642083), Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001), Nansha Key RD Program under Grant No.2022ZD014.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 3

[2] 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 *ICCV*, 2021. 2

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 2

[4] Alexandra Carlson, Manikandasriram S Ramanagopal, Nathan Tseng, Matthew Johnson-Roberson, Ram Vasudevan, and Katherine A Skinner. Cloner: Camera-lidar fusion for occupancy grid-aided neural representations. *IEEE Robotics and Automation Letters*, 2023. 5, 8

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 2

[6] Sayantan Datta, Carl Marshall, Zhao Dong, Zhengqin Li, and Derek Nowrouzezahrai. Efficient graphics representation with differentiable indirection. *arXiv:2309.08387*, 2023. 1

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022. 2, 3, 5

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, 2017. 3, 1

[9] Shuangkang Fang, Weixin Xu, Heng Wang, Yi Yang, Yufeng Wang, and Shuchang Zhou. One is all: Bridging the gap between neural radiance fields architectures with progressive volume distillation. In *AAAI*, 2023. 8

[10] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv:2203.15224*, 2022. 2, 1

[11] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv:2306.04988*, 2023. 2, 4, 6, 7

[12] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *TPAMI*, 2021. 1, 2

[13] 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 *ICCV*, 2023. 2

[14] Xiuzhong Hu, Guangming Xiong, Zheng Zang, Peng Jia, Yuxuan Han, and Junyi Ma. Pc-nerf: Parent-child neural radiance fields under partial sensor data loss in autonomous driving environments. *arXiv:2310.00874*, 2023. 2

[15] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. *ICCV*, 2023. 1, 2, 3, 6

[16] Xin Huang, Qi Zhang, Ying Feng, Xiaoyu Li, Xuan Wang, and Qing Wang. Local implicit ray function for generalizable radiance field representation. In *CVPR*, 2023. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 1

[18] Yixing Lao, Xiaogang Xu, Zhipeng Cai, Xihui Liu, and Hengshuang Zhao. Corresnerf: Image correspondence priors for neural radiance fields. In *NeurIPS*, 2023. 2, 3

[19] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *ICCV*, 2023. 8

[20] Zhuopeng Li, Lu Li, and Jianke Zhu. Read: Large-scale neural scene rendering for autonomous driving. In *AAAI*, 2023. 2

[21] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, 2023. 4

[22] Erich Liang, Kenan Deng, Xi Zhang, and Chun-Kai Wang. Hr-neus: Recovering high-frequency surface geometry via neural implicit surfaces. *arXiv:2302.06793*, 2023. 4

[23] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *TPAMI*, 2022. 2, 6

[24] Fan Lu, Yan Xu, Guang Chen, Hongsheng Li, Kwan-Yee Lin, and Changjun Jiang. Urban radiance field representation with deformable neural mesh primitives. In *ICCV*, 2023. 2

[25] 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 *CVPR*, 2023. 2

[26] 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*, 2021. 1, 2, 3, 6

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ToG*, 2022. 2, 3, 4, 6, 7

[28] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H Mueller, Chakravarty R Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *CGF*, 2021. 2, 3, 5

[29] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 1

[30] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022. 3

[31] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 2

[32] 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 *CVPR*, 2022. 2, 3

[33] Radu Alexandru Rosu and Sven Behnke. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *CVPR*, 2023. 8

[34] Andrew Sanders. *An introduction to Unreal engine 4*. AK Peters/CRC Press, 2016. 1

[35] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018. 1

[36] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2023. 8, 2, 3

[37] Claus Smitt, Michael Halstead, Patrick Zimmer, Thomas Läbe, Esra Guclu, Cyrill Stachniss, and Chris McCool. Pagnerf: Towards fast and efficient end-to-end panoptic 3d representations for agricultural robotics. *arXiv:2309.05339*, 2023. 8

[38] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 6

[39] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *SPL*, 2021. 3

[40] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022. 2

[41] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023. 1

[42] 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:2304.10406*, 2023. 1, 2, 3, 4, 6, 7

[43] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *CVPR*, 2023. 1, 2

[44] 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. *arXiv:2308.04413*, 2023. 2, 5

[45] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *CVPR*, 2023. 2

[46] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. 3

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[48] Guan Cheng Lee Wei Jong Yang. Addressing data misalignment in image-lidar fusion on point cloud segmentation. *arXiv:2309.14932*, 2023. 4

[49] Xinshuo Weng, Yunze Man, Jinhyung Park, Ye Yuan, Matthew O'Toole, and Kris M Kitani. All-in-one drive: A comprehensive perception dataset with high-density long-range point clouds. 2023. 2, 3, 6

[50] 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. *arXiv:2307.15058*, 2023. 1, 2

[51] Xiufeng Xie, Riccardo Gherardi, Zhihong Pan, and Stephen Huang. Hollownerf: Pruning hashgrid-based nerfs with trainable collision mitigation. In *ICCV*, 2023. 1

[52] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv:2303.00749*, 2023. 1, 2

[53] 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:2311.02077*, 2023. 1

[54] Yuanbo Yang, Yifei Yang, Hanlei Guo, Rong Xiong, Yue Wang, and Yiyi Liao. Urbangiraffe: Representing urban scenes as compositional generative neural feature fields. *arXiv:2303.14167*, 2023. 2

[55] 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 *CVPR*, 2023. 1, 2, 3, 4, 5, 6, 7

[56] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv:2112.05131*, 2021. 2

[57] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[58] Kaicheng Yu, Tang Tao, Hongwei Xie, Zhiwei Lin, Zhongwei Wu, Zhongyu Xia, Tingting Liang, Haiyang Sun, Jiong Deng, Dayang Hao, et al. Benchmarking the robustness of lidar-camera fusion for 3d object detection. *CVPRW*, 2022. 4

[59] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022. 2, 3

[60] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. *ICCV*, 2023. 1, 2, 3

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[62] Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Multi-task view synthesis with neural radiance fields. In *ICCV*, 2023. 2

[63] Zehan Zheng, Danni Wu, Ruisi Lu, Fan Lu, Guang Chen, and Changjun Jiang. Neuralpci: Spatio-temporal neural field for 3d point cloud multi-frame non-linear interpolation. In *CVPR*, 2023. 2

[64] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2