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

Supplementary Material

7. Limitations and Future work

As this paper is the first to reveal the misalignment issue in multimodal learning in NeRF, there remains room for improvement, which we would like to address in future work. As such, it is better suited to static scenes. Fortunately, there have been notable advancements in handling dynamic scenes [29, 50, 53, 55], and our decomposed encoding formulation can be seamlessly integrated with these advances. Moreover, for each dataset, it is necessary to search for the optimal alignment level of the coarse geometry. As indicated by the dynamic network technology [12], developing dynamic search levels holds the potential for further improvements in the alignment process. Furthermore, exploring more powerful fusion modules for alignment represents a promising research direction. Another impact of our coarse geometry alignment is the incorporation of multiple hash encoders, which introduce additional model parameters and computation. Nonetheless, we optimize per scene with two to three hours on a single NVIDIA GeForce RTX 3090 GPU, which is still much more cost-effective compared to traditional handcrafted game-engine-based virtual worlds [8, 34, 35], and there also have been efforts [6, 51] in improving the efficiency of hash encoders. Altogether, we hope that our work will inspire other researchers to contribute to the development of multimodal NeRF.

Discuss dynamic foreground objects. In Fig. 9, we illustrate the approach for dynamic scenes, where dynamic objects are modeled separately with the static background, and each object is transformed into its object-centroid coordinate system. This allows us to treat each object field as a small static scene, which can be directly extended to our AlignMiF. Additionally, dynamic objects pose more challenges, and our AlignMiF can provide fusion models with different levels of alignment for the dynamic objects and static background, providing a comprehensive solution.

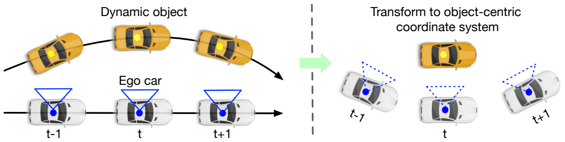

Figure 9. **The illustration of handling dynamic objects.**

8. Additional Details

Dataset. AIODrive consists of 100 video sequences generated by the CARLA Simulator, comprising about 100k labeled images and point cloud data. For our investigation, we utilize the provided mini-version of the dataset and masked the dynamic objects as Nerfstudio [41]. We select every 15th image in the sequences as the test set and take the remaining ones as the training set. KITTI-360 is a large-scale dataset containing over 320k images and 100k laser scans collected in urban environments with a driving distance of around 73.7 km. We select 4 static suburb sequences as PNF [10] and LiDAR-NeRF [42]. Each sequence contains 64 frames, with 4 equidistant frames for evaluation. For Waymo Open Dataset, we also select the 4 sequences mainly containing static objects for our experiments. We reserve every 10th frame as a test view and use the remaining about 188 samples for training.

Implementation details. Our AlignMiF is implemented based on open-source LiDAR-NeRF [42]. We optimize our AlignMiF model per scene with two to three hours of training time using a single NVIDIA GeForce RTX 3090 GPU. We use Adam [17] with a learning rate of 1e-2 to train our models. The coarse and fine networks are sampled 768 and 64 samples per ray, respectively. The finest resolution of the hash encoding is set to 32768. To better evaluate and compare the synthesis capability for details, we train and evaluate our methods and all the baselines with full-resolution images and LiDAR input. All ablation and analysis experiments were conducted on the sequence *seq-1908-1971* of KITTI-360 dataset and the segment *17761959194352517553_5448_420_5468_420* of Waymo dataset, which both are large scenes with numerous objects, making them ideal sequences for comparison.

9. Additional Results

Table 6. **Ablation study for the coarse geometry levels.**

Dataset	Levels	RGB Metric		LiDAR Metric	
		PSNR↑	SSIM↑	C-D↓	F-score↑
KITTI-360	4	23.68	0.773	0.080	0.929
	8	25.20	0.816	0.077	0.932
	16	24.49	0.803	0.091	0.921
Waymo	6	28.73	0.834	0.156	0.892
	9	29.22	0.841	0.151	0.896
	12	28.72	0.835	0.155	0.891

Ablations on the coarse geometry levels. As shown in Tab. 6, for each dataset, we search for the optimal alignment level, i.e., the β, of the coarse geometry. As indicated by

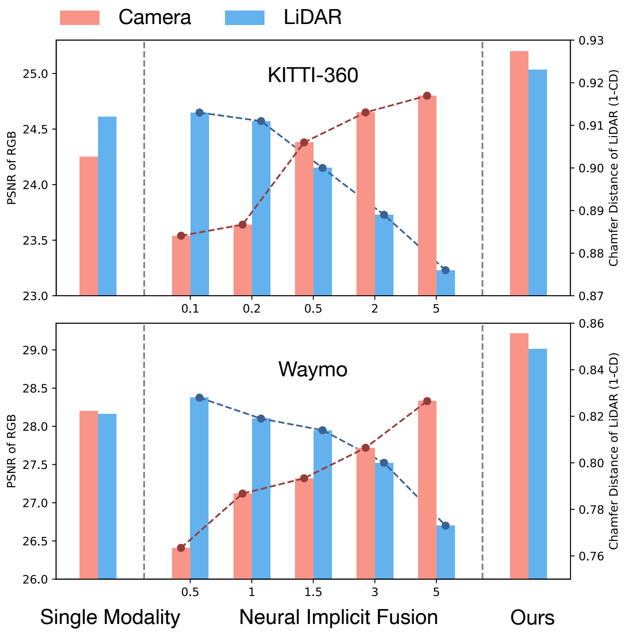

Figure 10. **The misalignment issue in multimodal implicit field.** For implicit neural fusion, there is a trade-off between the modalities due to the misalignment, making it challenging to improve both modalities simultaneously. Conversely, our method addresses the misalignment issue and achieves boosted multimodal performance. The horizontal axis denotes the weight ratio between the camera and LiDAR modality, λ_c/λ_l.

the dynamic network technology [12], developing dynamic search levels holds the potential for further improvements in the alignment process, and we left it as future work.

Misalignment issue in the multimodal implicit field. Previous multimodal implicit fields, e.g., UniSim [55], explored fusing multiple modalities within a single field, aiming to share implicit features from different modalities to enhance performance. However, the misaligned modalities often contradict each other, and as shown in Fig. 10, the line plot clearly illustrates a trade-off between the modalities: optimizing for one modality, such as the camera, can have a negative impact on the performance of another modality, such as LiDAR, and vice versa. In contrast, our AlignMiF effectively addresses the misalignment issue and achieves enhanced multimodal performance, as demonstrated by the bar chart in Fig. 10. We also provide qualitative visualization in Fig. 11.

Misalignment issue and different FOV. In Sec. 5.3, we attempted to detach the gradient of density from the camera modality to avoid geometry conflicts, i.e., *Detach RGB Density*, which is similar to gradients blocking in Panoptic-Lifting [36], but the FOV mismatch and misalignment issue have become significant obstacles as shown in Fig. 12. From (a)(b)(c), we can observe that the two modalities had different FOVs, resulting in the training being effective only in the pre-trained LiDAR FOV. Moreover, as demonstrated in Fig. 2 of Sec. 3.3 and (a)(c), LiDAR and camera exhibit variances in capturing finer details. Therefore, when solely relying on LiDAR for optimizing geometry, the resulting camera image and depth are both unsatisfactory, as illustrated by the distorted and thicker pole in (b) and (d). These figures further emphasize the importance of addressing the misalignment issue.

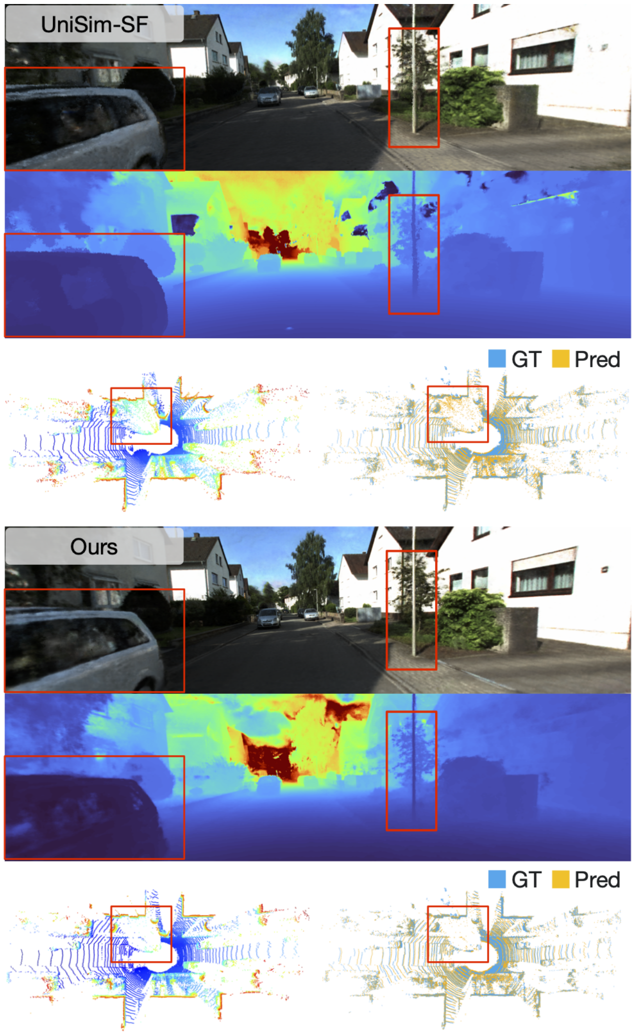

Figure 11. **The misalignment issue in multimodal implicit field.** Implicit fusing misaligned modalities leads to suboptimal results, i.e., the blurred image and messy points, as depicted in the figure. Our AlignMiF enhances the alignment and fusion between LiDAR and camera modalities, leading to more accurate join synthesis of novel views (zoom-in for the best of views).

State-of-the-art results on AIODrive dataset. As shown in Tab. 7, our AlignMiF, achieves superior results even on the AIODrive synthetic dataset without the misalign-

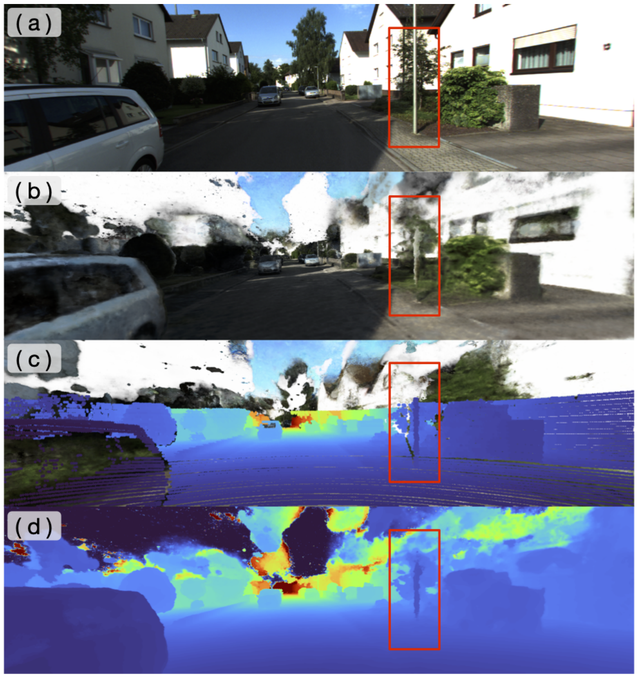

Figure 12. **Misalignment issue and different FOV of LiDAR and camera.** (a) Original image, (b) Rendered image, (c) Rendered image with projected points from associate LiDAR frame, (d) Rendered image depth.

Table 7. **State-of-the-art results on AIODrive dataset.**

Method	M	RGB Metric		LiDAR Metric	
		PSNR↑	SSIM↑	C-D↓	F-score↑
i-NGP [27]	C	34.43	0.893	–	–
LiDAR-NeRF [42]	L	–	–	0.178	0.873
UniSim-SF [55]	LC	34.53 (+)	0.904	0.153 (+)	0.905
AlignMiF (Ours)	LC	**34.82 (+)**	0.908	**0.123 (+)**	0.915

M, L, C denotes modality, LiDAR, camera respectively.

ment issue, outperforming the previous implicit fusion approach. This can be attributed to the fact that despite having the same underlying scene geometry, the required features for different modalities and representations might differ slightly, such as the LiDAR intensity and image color. This also aligns with the results of *Share Coarse-Geo* in Tab. 5 of Sec. 5.3 and the observation in Panoptic-Lifting [36]. These results both demonstrate the efficiency of our network design.

Details results on KITTI-360 and Waymo datasets. We report detailed results on the sequences of the KITTI-360 and Waymo datasets in Tab. 10. Our AlignMiF consistently outperforms the baselines over all sequences in all metrics. The details of sequences are also shown in Tab. 10.

Table 8. **Enhancing the detection model with AlignMiF.**

Method	L1_mAP	L1_mAPH	L2_mAP	L2_mAPH
TransFusion [1]	38.71	35.05	33.98	30.79
+ AlignMiF	40.18 **(+1.47)**	36.39**(+1.34)**	35.31**(+1.33)**	32.01 **(+1.22)**

Table 9. **Computational cost (microsecond) for rendering 4096 rays.**

Method	Hash-Encoding	Geo-MLP	Color-MLP
UniSim-SF [55]	86	72	96
AlignMiF	86 (SGI) + 156 (GAA)	72	96

Boosting downstream applications. We are eager to explore the potential benefits of improving downstream applications. We choose the powerful LiDAR-camera fusion detection method TransFusion [1] and employ our AlignMiF to generate more diverse sensor data for data augmentation. Due to computational constraints, we conducted experiments on a limited number of Waymo scenes. The results in Tab. 8 demonstrate the effectiveness of our approach in enhancing the performance of the downstream model.

Computational complexities. In Tab. 9, we present the computational cost to facilitate further research.

9.1. Qualitative Results

Qualitative results on KITTI-360 dataset. We provide more qualitative results on KITTI-360 dataset in Fig. 13 and Fig. 15, which show the mutual benefits of our AlignMiF. The LiDAR modality significantly improves the learning of image and depth quality in the camera, while the semantic information from RGB assists the LiDAR in better converging to object boundaries.

Qualitative results on Waymo dataset. We provide more qualitative results on the Waymo dataset in Fig. 14 and Fig. 15, which demonstrate that the proposed AlignMiF significantly enhances the alignment and fusion between LiDAR and camera modalities, leading to more accurate join synthesis of LiDAR and camera novel views.

Video demo. In addition to the figures, we have attached a video demo in the supplementary materials, which consists of hundreds of frames that provide a more comprehensive evaluation of our proposed approach.

Table 10. **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↓
Sequence		*Seq 1538–1601*						*seg11379226583756500423_6230_810_6250_810*					
i-NGP [27]	C	25.22	0.831	0.175	–	–	–	29.26	0.825	0.369	–	–	–
LiDAR-NeRF [42]	L	–	–	–	0.088	0.925	0.106	–	–	–	0.216	0.853	0.026
UniSim-SF [55]	LC	22.92 (−)	0.746	0.328	0.083 (+)	0.927	0.103	26.90 (−)	0.777	0.399	0.199 (+)	0.854	0.027
UniSim-SF [55]$^{\triangledown}$	LC	25.25 (+)	0.827	0.184	0.109 (−)	0.904	0.102	29.35 (+)	0.825	0.367	0.359 (−)	0.761	0.030
AlignMiF	LC	25.67 (+)	0.837	0.176	0.079 (+)	0.930	0.106	30.16 (+)	0.838	0.331	0.191 (+)	0.863	0.026
Sequence		*Seq 1728–1791*						*seg10676267326664322837_311_180_331_180*					
i-NGP [27]	C	24.90	0.816	0.167	–	–	–	29.52	0.861	0.299	–	–	–
LiDAR-NeRF [42]	L	–	–	–	0.107	0.895	0.111	–	–	–	0.264	0.841	0.026
UniSim-SF [55]$^{\triangle}$	LC	23.86 (−)	0.782	0.228	0.097 (+)	0.909	0.094	26.47 (−)	0.808	0.341	0.254 (+)	0.848	0.026
UniSim-SF [55]$^{\triangledown}$	LC	25.38 (+)	0.823	0.167	0.127 (−)	0.891	0.093	29.54 (+)	0.862	0.297	0.612 (−)	0.704	0.039
AlignMiF	LC	25.43 (+)	0.836	0.148	0.086 (+)	0.913	0.096	30.27 (+)	0.873	0.273	0.228 (+)	0.859	0.025
Sequence		*Seq 1908–1971*						*seg17761959194352517553_5448_420_5468_420*					
i-NGP [27]	C	24.45	0.787	0.184	–	–	–	28.20	0.830	0.372	–	–	–
LiDAR-NeRF [42]	L	–	–	–	0.088	0.920	0.159	–	–	–	0.179	0.885	0.049
UniSim-SF [55]$^{\triangle}$	LC	23.54 (−)	0.759	0.235	0.087 (+)	0.929	0.097	26.41 (−)	0.789	0.403	0.173 (+)	0.891	0.049
UniSim-SF [55]$^{\triangledown}$	LC	24.65 (+)	0.803	0.172	0.111 (−)	0.912	0.097	28.33 (+)	0.830	0.369	0.227 (−)	0.840	0.052
AlignMiF	LC	25.20 (+)	0.816	0.160	0.077 (+)	0.932	0.101	29.22 (+)	0.841	0.327	0.151 (+)	0.896	0.048
Sequence		*Seq 3353–3416*						*seg11724067803607999916_1660_000_1680_000*					
i-NGP [27]	C	23.88	0.800	0.199	–	–	–	28.30	0.810	0.480	–	–	–
LiDAR-NeRF [42]	L	–	–	–	0.094	0.927	0.112	–	–	–	0.127	0.908	0.057
UniSim-SF [55]$^{\triangle}$	LC	22.90 (−)	0.746	0.283	0.091 (+)	0.933	0.093	26.91 (−)	0.778	0.526	0.118 (+)	0.919	0.056
UniSim-SF [55]$^{\triangledown}$	LC	24.51 (+)	0.797	0.213	0.110 (−)	0.919	0.091	28.72 (+)	0.816	0.463	0.225 (−)	0.841	0.058
AlignMiF	LC	24.91 (+)	0.815	0.175	0.082 (+)	0.937	0.096	29.47 (+)	0.828	0.427	0.107 (+)	0.921	0.054
Average								Average					
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	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.

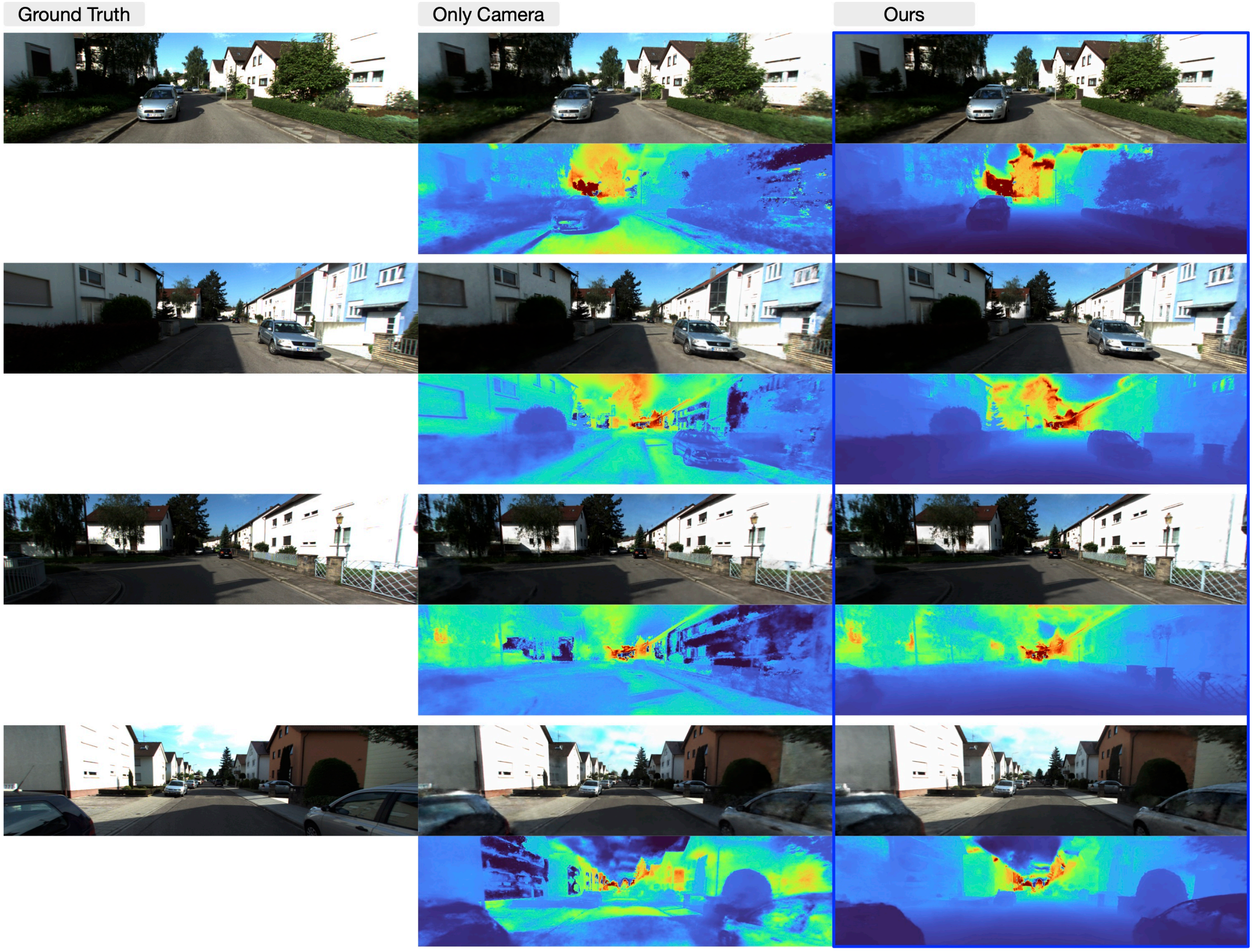

Figure 13. **Qualitative results of the camera on KITTI-360 dataset.** Our AlignMiF enhances information interactions between modalities and improves image and depth quality in the camera using LiDAR information.

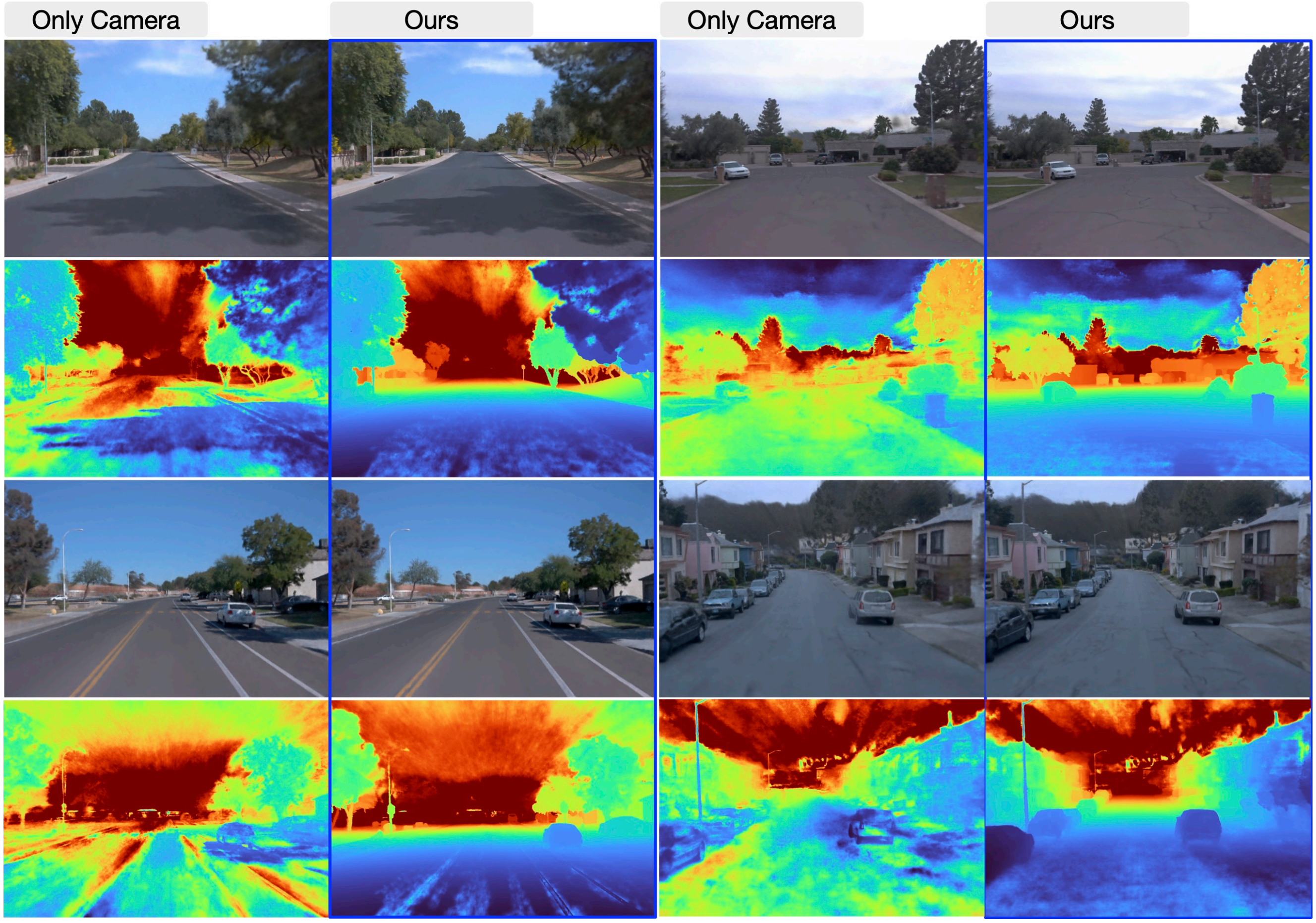

Figure 14. **Qualitative results of the camera on Waymo dataset.** Our AlignMiF enhances information interactions between modalities and improves image and depth quality in the camera using LiDAR information.

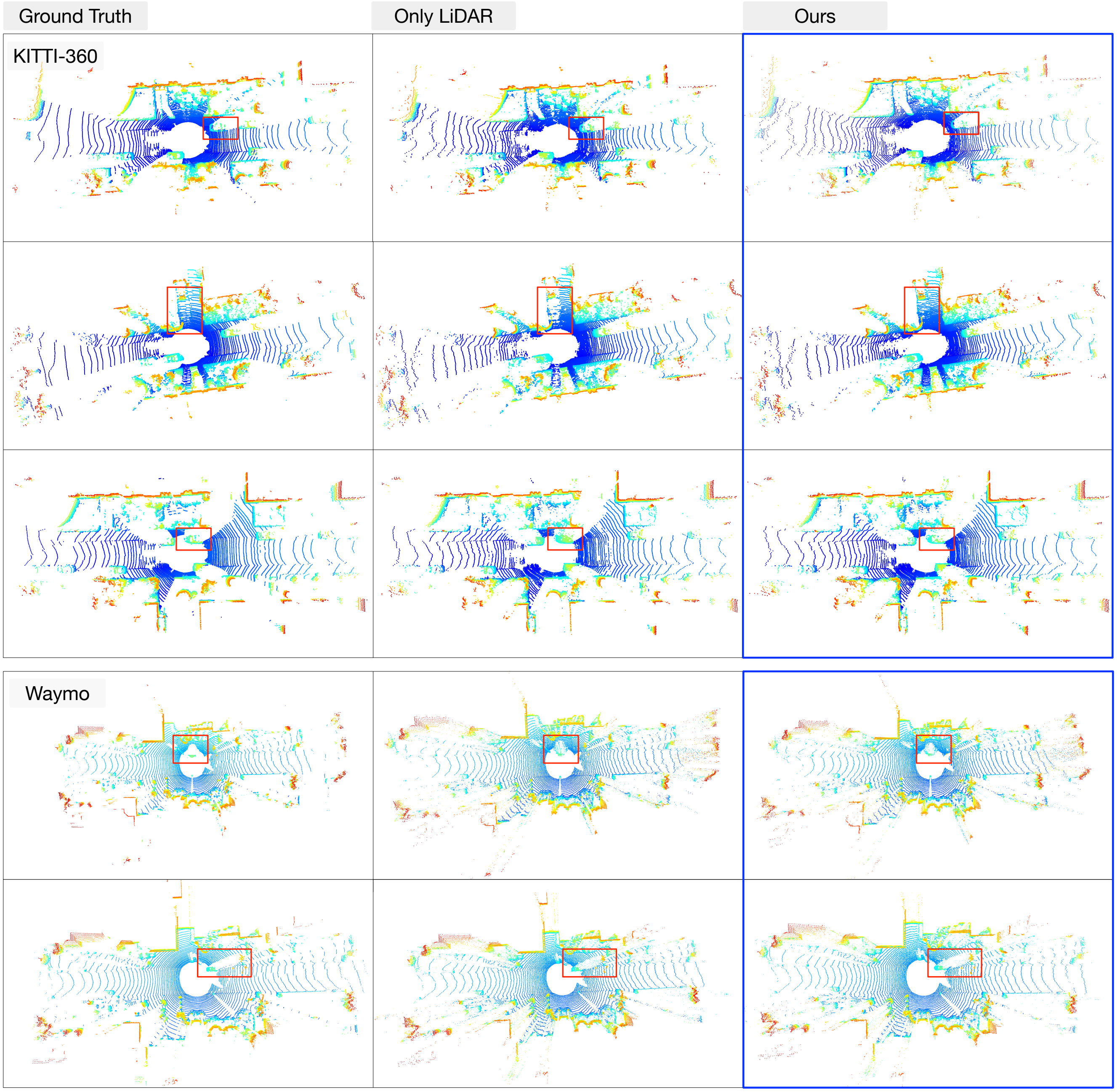

Figure 15. **Qualitative results of the LiDAR on KITTI-360 and Waymo datasets.** Our AlignMiF enhances information interactions between modalities and semantic information from the camera aids the LiDAR in better converging to object boundaries (zoom-in for the best of views). Visualizing from a single perspective may not provide a comprehensive analysis of the LiDAR in 3D space. It's encouraged to try our code and models, and use 3D-view tools for a more comprehensive understanding of our method's superiority.