Supplementary Material for: Learning to Fuse Monocular and Multi-view Cues for Multi-frame Depth Estimation in Dynamic Scenes

Rui Li[1], Dong Gong[2], Wei Yin[3], Hao Chen[4], Yu Zhu[1], Kaixuan Wang[3],
Xiaozhi Chen[3], Jinqiu Sun[1], Yanning Zhang[1]
[1]Northwestern Polytechnical University, [2]The University of New South Wales, [3]DJI, [4]Zhejiang University

1. Identifying Dynamic Areas in DDAD

Since there are no previously defined dynamic masks in the DDAD dataset [8], we use a simple and effective strategy to identify the dynamic masks for evaluation (Sec.5.1 of the main paper), which differs from [11] in using the GT depth with simpler steps. The procedures are as follows:

1) **Instance segmentation.** We use the off-the-shelf instance segmentation method [2] in MMDetection [3] to segment possible movable instances on every image. The movable instances include 'car', 'truck', 'trailer', 'bus', 'construction vehicle', 'bicycle', 'motorcycle' as well as 'pedestrian'. We filter out the instance masks that are smaller than 1000 pixels or have confidence scores lower than 0.8.

2) **Warping adjacent images to the target view**. We use the ground truth depth $\hat{D}$, the known camera extrinsic T as well as the intrinsic K to warp adjacent images I_{t-1}, I_{t+1} to the target view t. Since the GT depth is sparse, we conduct depth completion to the GT depth maps using the dilation operation. The warped images $I_{t-1\to t}, I_{t+1\to t}$ are computed following operations in [6, 11]. We then compute the SSIM-based [9] photometric error between the warped images and the target image

$$\begin{aligned} pe(I_{t-1\to t}, I_t) &= \text{SSIM}(I_{t-1\to t}, I_t), \\ pe(I_{t+1\to t}, I_t) &= \text{SSIM}(I_{t+1\to t}, I_t). \end{aligned} \tag{1}$$

3) **Thresholding photometric error to identify dynamic areas.** For each segmented instance mask, we compute its average photometric error from $pe(I_{t-1\to t}, I_t)$ and $pe(I_{t+1\to t}, I_t)$. Instance masks with photometric error larger than 0.15 are regarded as dynamic objects.

We visualize the intermediate results as well as the final identified dynamic mask in Fig. 1. The used method successfully distinguishes dynamic objects from the surrounding environment as well as the static object instances.

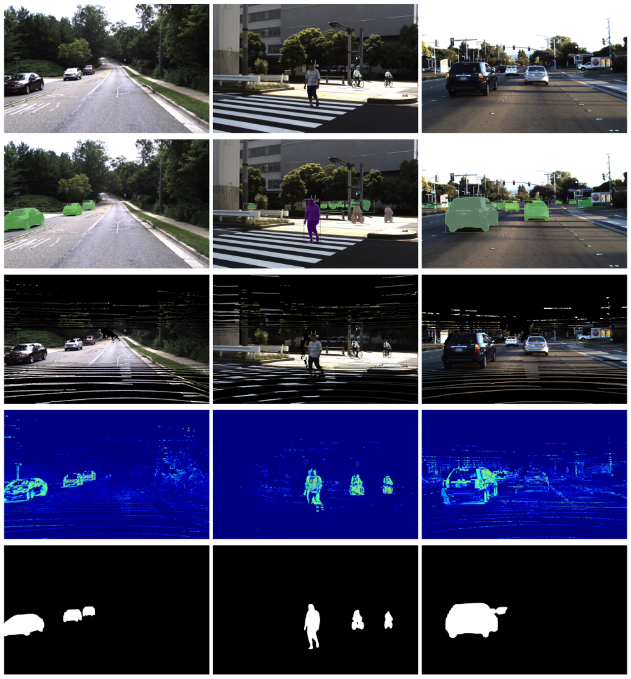
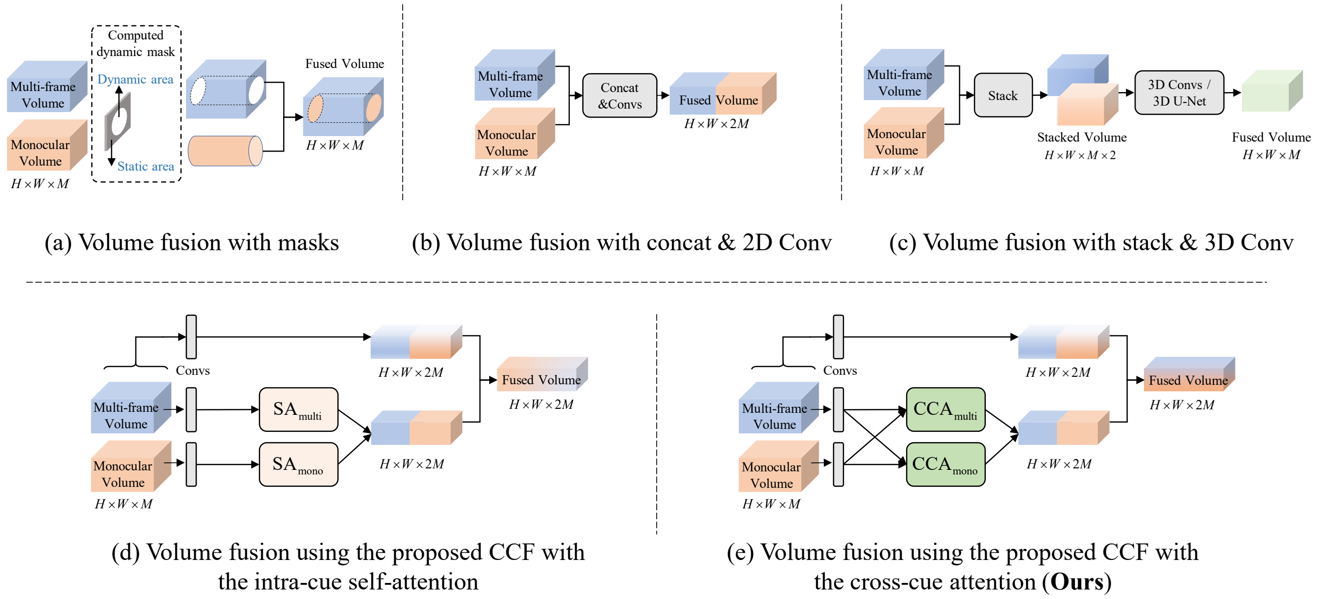


Figure 1. **Intermediate results during dynamic mask identification.** From top to bottom: the target image, instance segmentation masks (before filtering), the warped adjacent image, the photometric error map, and the final dynamic area masks. The used strategy managed to distinguish dynamic areas from surrounding environments (column #1~2) as well as other static instances (column #3).

2. Network Details in the Analysis Experiment

In Sec. 3 of the main paper, we evaluate the behaviors of multi-view and monocular cues by experimenting on pure multi-frame and monocular networks. For the pure multi-frame network, we use the U-Net architecture introduced in [11], which takes the cost volume as well as the image context to regress multi-frame depth. The depth network is supervised using ground truth depth. We do not use any masking strategies or bootstrapping training schemes [11] for the multi-frame network.



(a) Volume fusion with masks

(b) Volume fusion with concat & 2D Conv

(c) Volume fusion with stack & 3D Conv

(d) Volume fusion using the proposed CCF with the intra-cue self-attention

(e) Volume fusion using the proposed CCF with the cross-cue attention (**Ours**)

Figure 2. **Network details in ablation study.** We show different versions (a)~2 (d) of our volume fusion scheme as discussed in Sec.5.4 of the main paper. We also show a simplified diagram of our method (e) for comparison.

For the monocular network, we use a U-Net architecture similar to that of [11], with the difference that we remove the convolutions for the input cost volume, which is absent in the monocular depth estimation task. The monocular network is also trained with ground truth depth.

The similar architectures between multi-frame and monocular networks ensure that the observed different behaviors of the two networks come from the inherent properties of multi-view and monocular cues, rather than the network architectures.

3. Network Details in Ablation Study

We introduce in detail and show diagrams of the different variants of our method as initially discussed in Sec. 5.4 of the main paper.

Volume fusion with explicit masks. As shown in Fig. 2 (a), given both multi-frame and monocular volumes, we fuse the two volumes using masks by first masking out the dynamic areas (the white area) in the multi-frame volume then compensating the dynamic areas with the monocular volume. The fused volume is then sent to the depth module for final depth prediction.

Volume fusion with concatenation & 2D convs. As shown in Fig. 2 (b), we directly concatenate the multi-frame and monocular volumes in the channel dimension, yielding a volume with size $(H \times W \times 2M)$. We then use 3 convolution layers to process this volume, yielding the fused volume that is then sent to the depth network. This strategy is named ‘Plain Volume Fusion’ in Sec.3 of the main paper.

Volume fusion with stack & 3D convs. As shown in Fig. 2 (c), we stack the multi-frame and monocular volume in a new dimension, yielding a 4D tensor of size $(H \times W \times M \times 2)$. We then use 3D convolutions to process the tensor with either 3D convolution layers or a 3D convolution U-Net, following the practices of the MVS methods [7, 12].

Our CCF with intra-cue self-attention. We fuse the multi-frame and monocular volumes under the proposed *cross-cue* fusion (CCF) module but substitute the proposed cross-cue attention (CCA) with an intra-cue self-attention (SA). As shown in Fig. 2 (d), our CCF with intra-cue self-attention (SA) mainly differs from the CCF with cross-cue attention (CA) in the input end as well as the way to compute the attention maps.

4. Analysis on Computation Complexity

As shown in Table 1, with reasonably more computation, the proposed method significantly outperforms the baseline (using only multi-frame cues) and MonoRec [11]. It can outperform MaGNet [1] on depth accuracy with much less computation.

Method	Baseline	MonoRec [11]	MaGNet [1]	**Ours**
Params (M)	16.15	17.65	76.38	20.67
FLOPs (G)	64.65	93.78	263.92	129.31
Abs Rel (Dynamic)	0.382	0.360	0.141	**0.118**

Table 1. Comparison of computation complexity and performance.

5. More Qualitative Results

We show more qualitative comparisons in KITTI [5] (in Fig.3), as well as the comparisons on cross-dataset performance in DDAD [8] (in Fig.5). We can observe that our method predicts more reasonable scene geometry than other methods, especially in dynamic areas.

Besides the qualitative comparisons, we also provide more visualization results of our method as shown in Fig. 4 and Fig. 6. Our method predicts reasonably accurate depth maps in both static and dynamic areas of the scene.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2842–2851, 2022. 2, 4, 6

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 1

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[4] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. *arXiv preprint arXiv:2203.15174*, 2022. 4

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 3, 4, 5

[6] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019. 1

[7] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2

[8] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 1, 3, 6

[9] 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. 1

[10] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 4

[11] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6112–6122, 2021. 1, 2, 4, 6

[12] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

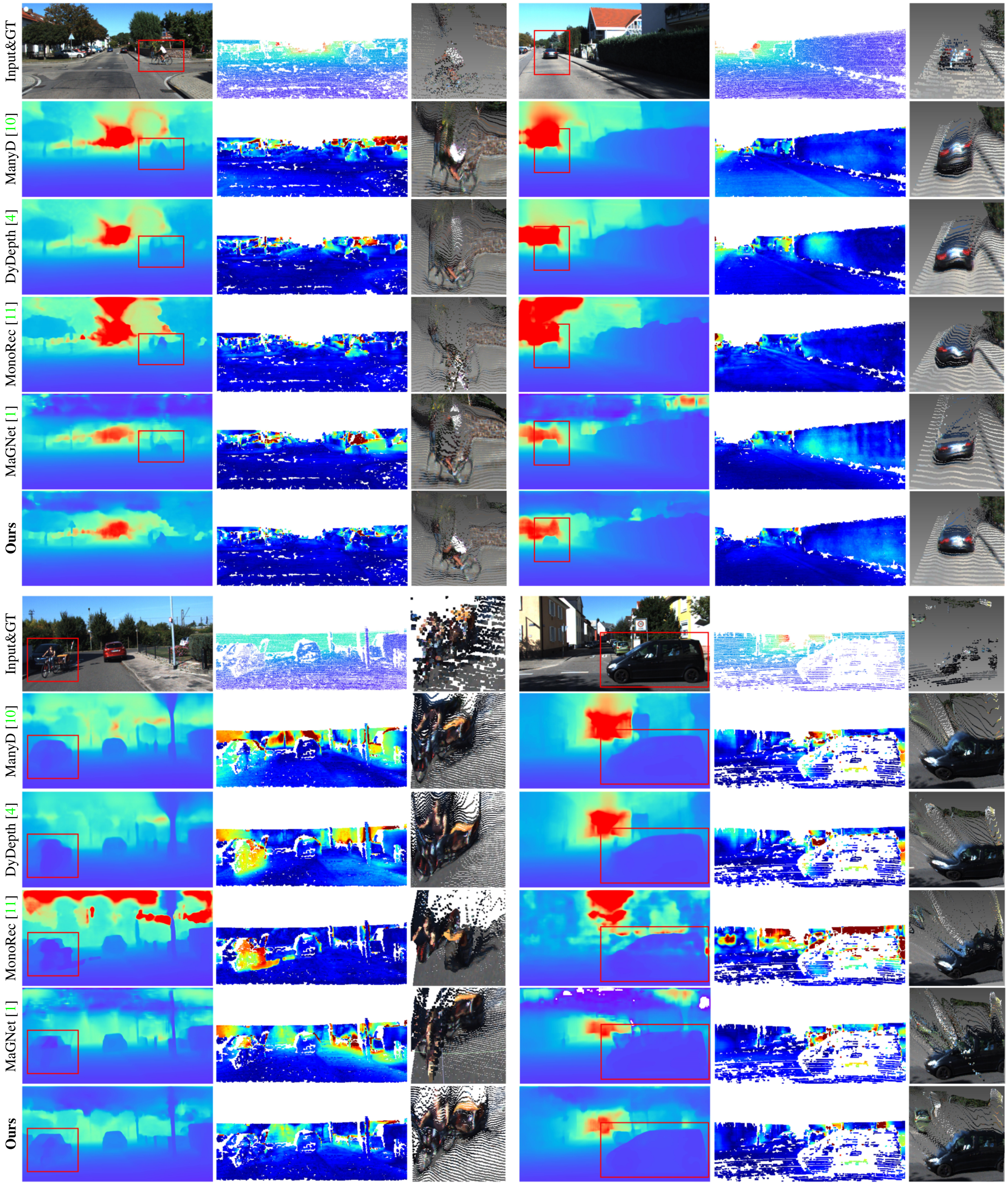


Figure 3. **Qualitative comparisons on KITTI dataset [5].** From left to right: depth predictions (dynamic objects are highlighted with red boxes), error maps, and the reconstructed point clouds of dynamic areas. Our method achieves the best dynamic results and reconstructs more reasonable object shapes than state-of-the-art methods.

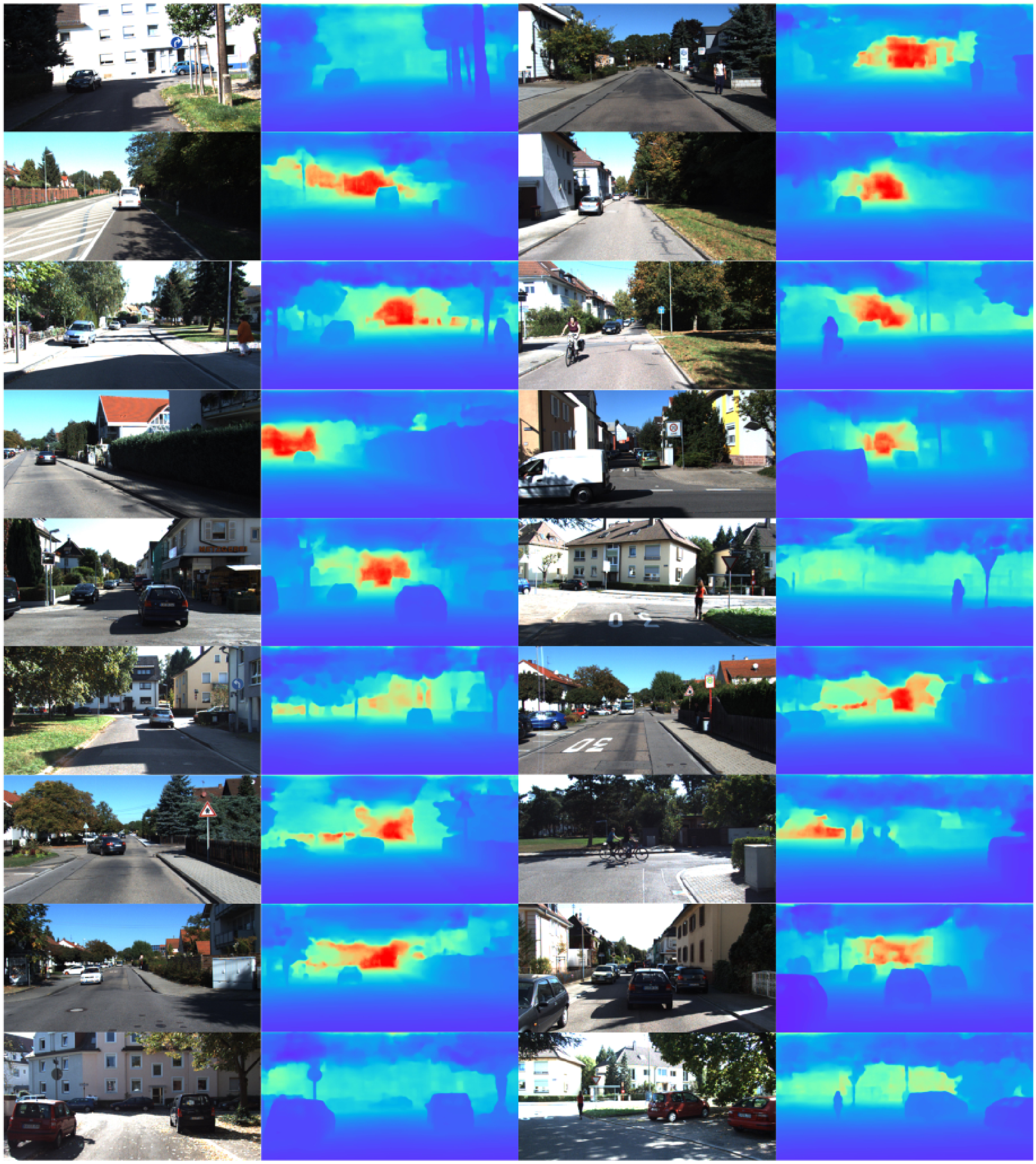


Figure 4. **Qualitative results of our method on KITTI [5].** We provide more qualitative results to validate our method's effectiveness for depth estimation in dynamic scenes. Our method conducts accurate scene depth estimations with sharp object boundaries as well as reasonable dynamic object shapes.

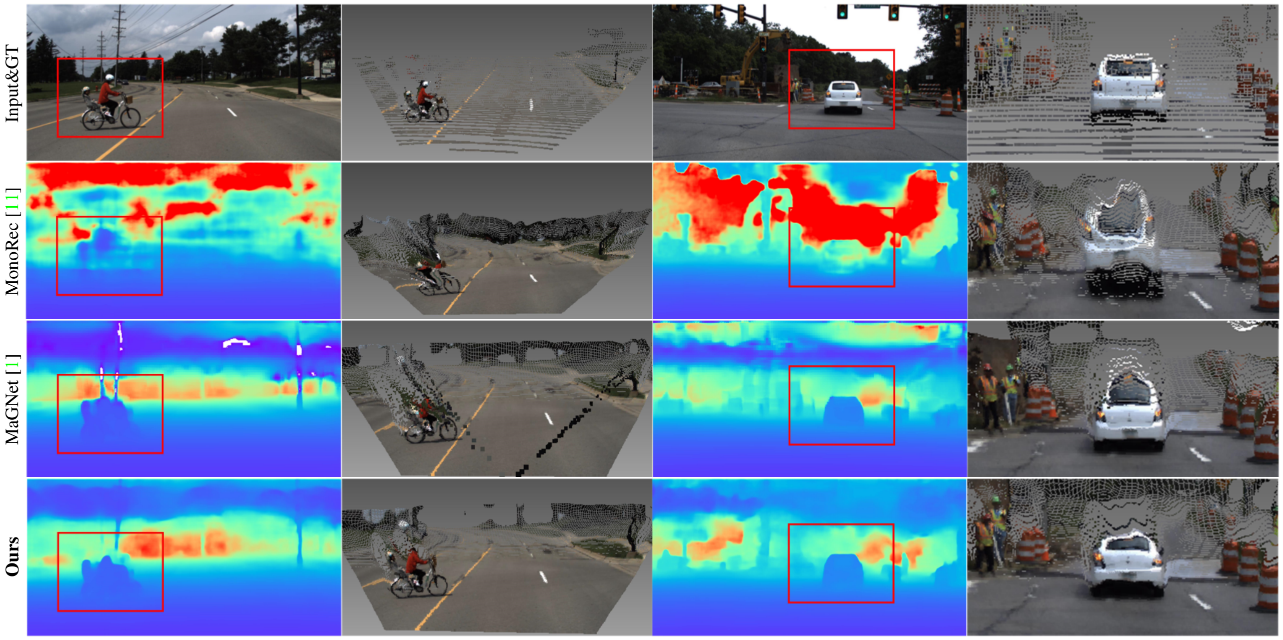


Figure 5. **Qualitative comparisons between KITTI-trained models on DDAD dataset [8].** . From left to right: input image & depth predictions (dynamic objects are highlighted with red boxes), the reconstructed point clouds of the scene. Our method better reconstructs the scene than other methods, especially in dynamic areas.

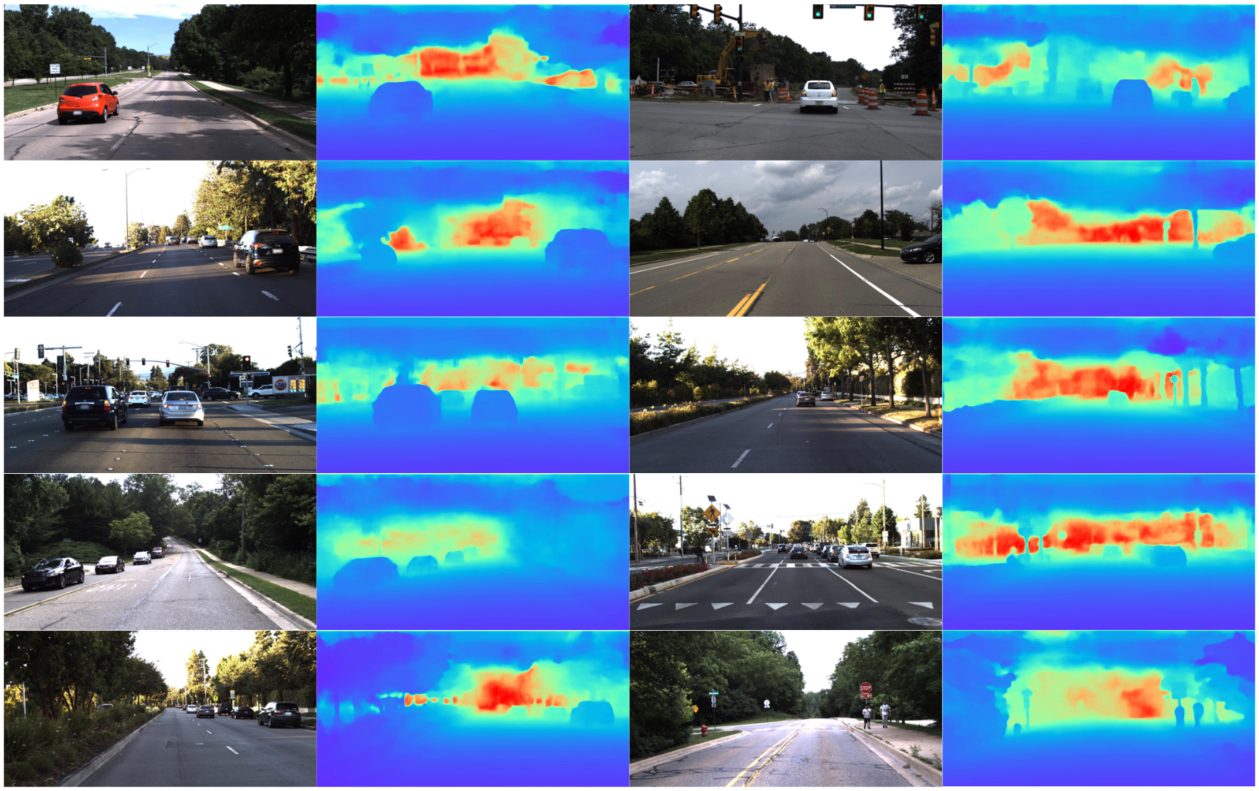


Figure 6. **Visualization of our method's cross-dataset results on DDAD dataset [8].** We provide more visualization results of our method to evaluate its cross-dataset depth estimation performance. Our method generates reasonable scene structures of the whole scene including dynamic areas.