RoSAMDepth: Robust Self-supervised Depth Estimation Leveraging Segment Anything Model

Supplementary Material

1. Methods Details

1.1. SAM Mask Generation

SAM often produces a hierarchical set of masks for a single object, detecting not only the entire object but also its constituent parts (e.g., wheels, doors, windows, lights, and even bumpers for a car, as shown in Fig. 1). While detailed, this hierarchy introduces redundancy and potential noise for operations requiring object-level information. To address this, we adopt a simple yet effective strategy. We assume that in overlapping masks, the mask with the largest spatial extent represents the object instance, while smaller nested masks represent its parts.

Formally, let $M_{sam}^{raw} \in \{0,1\}^{K\times H\times W}$ denote the set of K binary masks produced by SAM for an input image I, where $M_{sam,i}^{raw} \in \{0,1\}^{H\times W}$ is the i-th mask. Specifically, we iterate through the masks with an area below a threshold (set to 10^5 pixels, approx. 200×500). If a small mask is significantly covered by a larger mask (overlap ratio > 0.75), it is considered a sub-part and is merged into the larger parent mask. As shown in Fig. 1, the resulting unified segmentations serve as robust object-level priors, ensuring that the network achieves consistent and accurate depth predictions across the entire object instance.

1.2. Training Framework

We build upon the Syn2Real-Depth framework [6] to investigate how object-level priors from SAM [3] can enhance robust self-supervised depth estimation in diverse real-world scenes. Syn2Real-Depth serves as an ideal baseline due to its real adaptation stage designed for diverse real data. The framework employs ManyDepth [5] as the depth backbone, originally defining three domain-specific models: Φ_{day} (trained on real daytime data), Φ_{syn} (trained on diverse synthetic data), and Φ_{real} (trained on diverse real data). The pose network follows the standard architecture described in [2, 5]. Since our contribution focuses on integrating SAM-based guidance into the real adaptation stage, we simplify the notation to better reflect our training paradigm. Specifically, we designate Φ_{syn} as the initial teacher Φ_t, and Φ_{real} as the student Φ_s, which is trained on real data with object-level priors.

1.3. Implementation Details

Our model is trained on a single NVIDIA 4090D GPU, taking about 10 hours. Regarding the hyperparameters, we set the temperature $T = 1.0$ and the number of feature scales $S = 3$ for L_{src}. For the temperature-controlled sharpening sigmoid function $S_\kappa(\cdot)$, the sharpening temperature is set to $\kappa = 15.0$. In the object-level reliability estimation (ORE) strategy, we employ a scaling factor $\beta = 1.0$ and a bias term $\epsilon = 1.0$. Finally, the balancing weights for the total objective L_{total} are set to $\lambda_1 = 0.05$ and $\lambda_2 = 0.2$.

1.4. Supervision Details

To ensure robust performance under adverse conditions, our training on real data inherits the auxiliary losses from Syn2Real-Depth [6]. Specifically, we incorporate the external loss term L_{ext}, which comprises the K-L divergence (L_{dis}^{KL}), the cost volume learning loss (L_{cv}), and the supervision for pose (L_T), following the exact formulation in [6]. We retain these components unchanged to leverage their proven efficacy in the real data. This allows us to isolate the source of performance gains, demonstrating that our proposed object-level integration provides additive improvements over a strong baseline.

2. More Qualitative Examples

As supplementary to the main paper, we provide more qualitative results on nuScenes [1] (day-clear, night, day-rain), Robotcar [4] (daytime, nighttime), as shown in Figs. 2 and 3.

Figure 1. Example of the SAM masks generation.

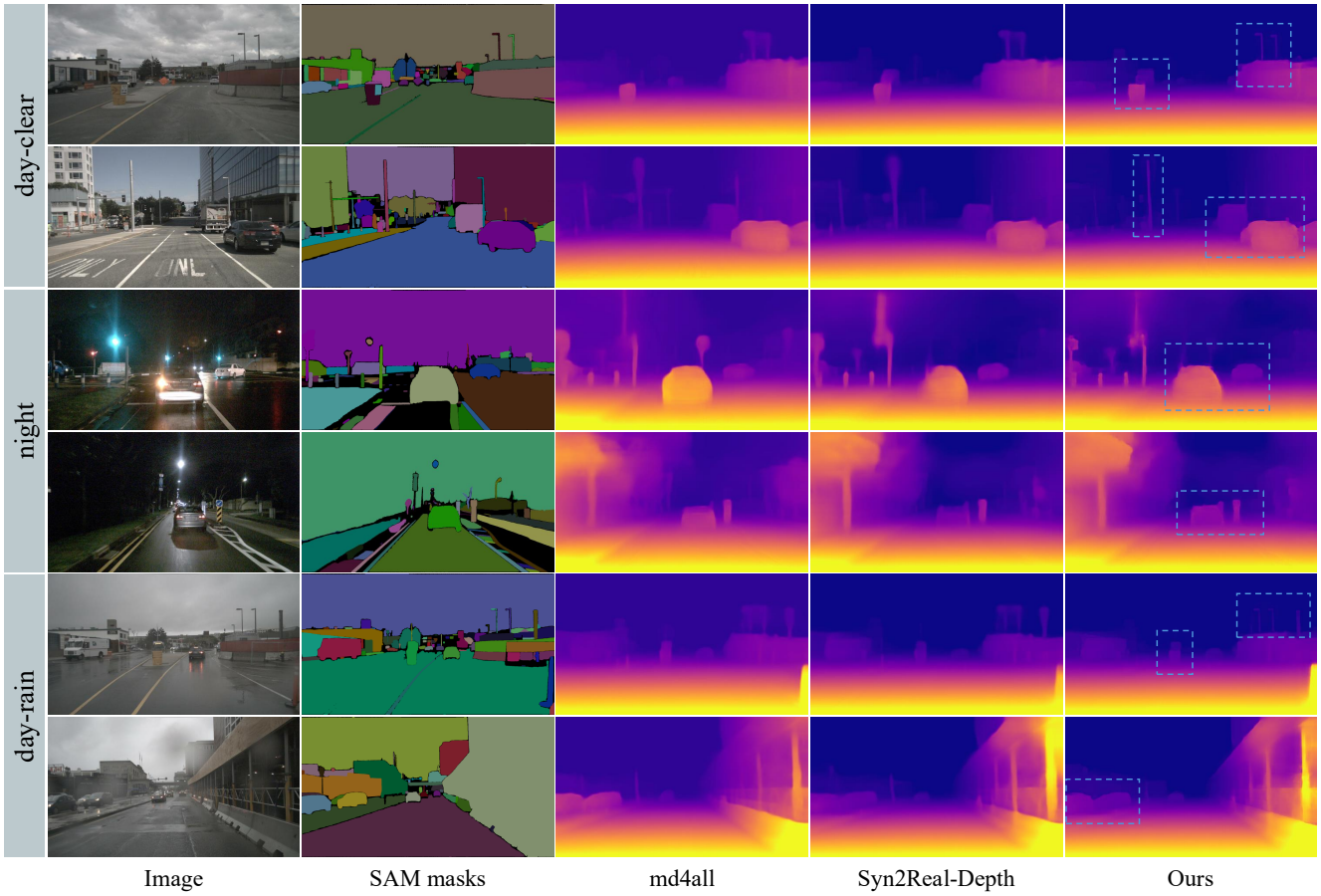

Figure 2. More qualitative results on nuScenes [1]

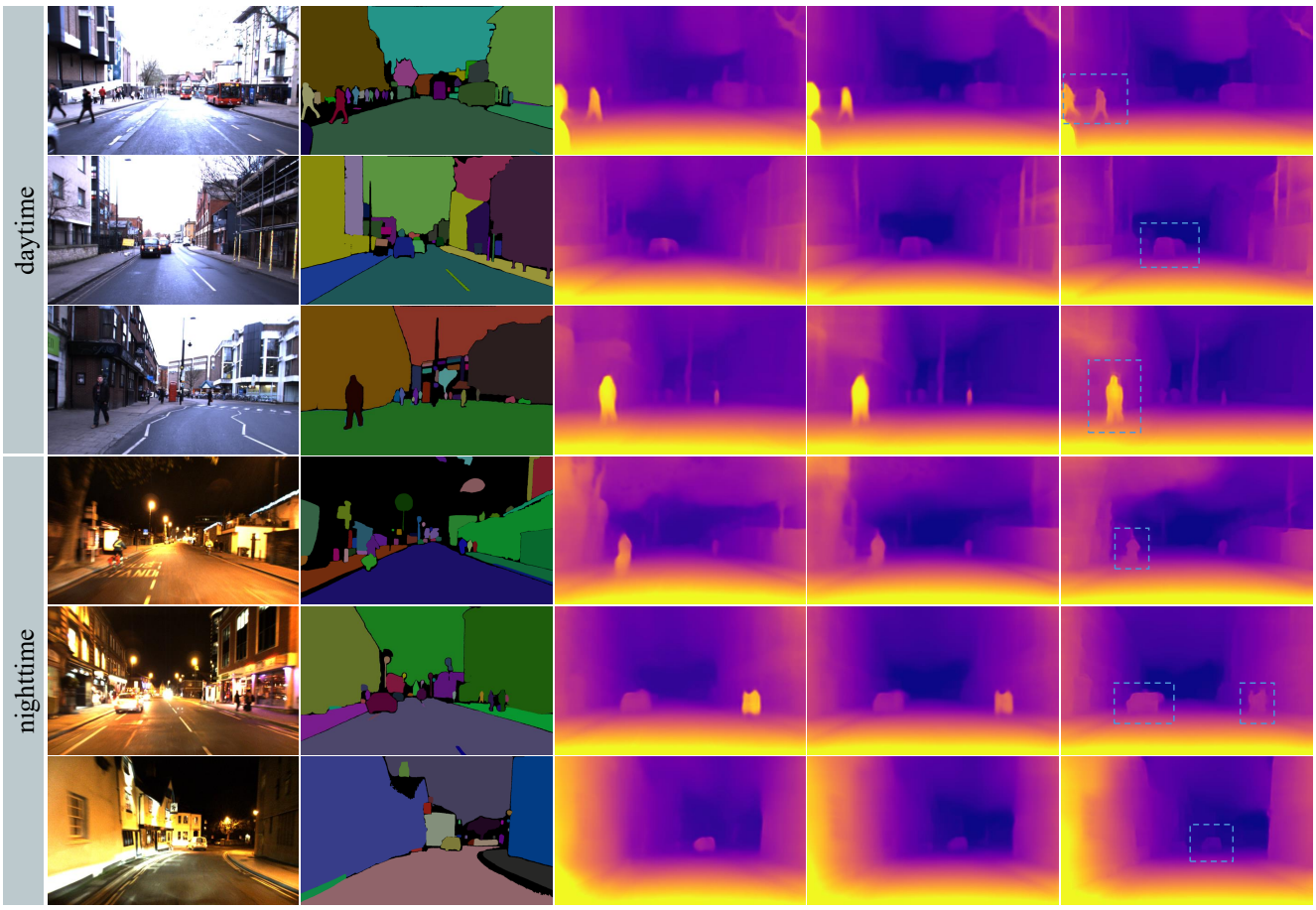

Figure 3. More qualitative results on Robotcar [4]

References

[1] 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. 1, 2

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[4] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 1, 3

[5] 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. 1

[6] Weilong Yan, Ming Li, Haipeng Li, Shuwei Shao, and Robby T Tan. Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21880–21890, 2025. 1