Monocular Depth Estimation 2025: Segmentation-Guided Refinement for Depth Estimation

Yuto Kageyama Riku Itsuji Rintaro Otsubo Kanta Sawafuji
Takuya Kurihara Tomoki Abe Hideo Saito
Keio University
Yokohama, Kanagawa, Japan
{y-kage.shadow96, diosa.de.gato.359, rintarootsubo, sawafuji_kanta}@keio.jp
{takuyakurihara040130, abe.tomoki.1106, hs}@keio.jp

Abstract

This paper presents our solution to the Monocular Depth Estimation 2025 Challenge[3]. Depth estimation is essential for various applications, including autonomous driving, augmented reality, robotics, and 3D reconstruction. In our approach, we incorporate segmentation masks to highlight regions that are difficult to predict and to leverage prior structural knowledge. These masks are generated in advance using a state-of-the-art segmentation model.

1. Analysis

1.1. Baseline Analysis

We adopt Depth Anything V2[8] as our baseline model. It achieves reasonably good performance in general, but there are two main issues. First, depth estimation around people tends to be inaccurate, often resulting in unnatural depth artifacts. Second, although the soccer field should ideally appear smooth, the reconstructed depth maps frequently exhibit undesired unevenness. These observations suggest that additional structural priors or localized enhancements are necessary for a better depth prediction in this domain.

1.2. Dataset Analysis

The dataset consists of high-resolution soccer video frames (1920×1080). The dataset resolution is significantly higher compared to the typical input resolutions used for pre-trained models such as ImageNet backbones (e.g., 224×224 or 384×384). Therefore, we fine-tune the baseline models to better adapt to the large input size. On the other hand, the videos' temporal resolution is relatively low, leading to significant scene changes even between adjacent frames. This characteristic makes them unsuitable for video-based methods that rely on temporal continuity, so we focus on single-frame depth estimation instead.

We also analyze ground-truth depth maps by converting them into point clouds. Regardless of the assumed focal length, the point clouds consistently exhibit curvature along the z-axis. This suggests that the soccer field, which should ideally be flat, is not perfectly flat even in the ground-truth data. Based on this observation, we apply a post-processing step that smooths the predicted depth maps while preserving the overall field geometry, without enforcing perfect flatness.

2. Method

Our approach improves the baseline in two key ways based on the observations in Section 2.

First, to address inaccurate depth estimation around players, we generate player masks using Grounded SAM2[1, 2, 4–7] with the prompt *"person."*. During training, we apply a higher weight to the loss computed over these masked regions, encouraging the model to focus more on the players and improve their depth accuracy.

We also generate field masks using Grounded SAM2 with the prompt *"grass."*. After depth estimation, we apply a post-processing step where we use median and average filters, but only consider valid pixels inside the field mask. This approach helps mitigate the impact of outliers and prevents unnecessary smoothing of edges, while maintaining local consistency within the field region.

Overall, our method combines region-specific supervision and structural priors to guide the model toward more accurate and realistic depth estimation, without relying on temporal information between frames.

3. Results

We fine-tune the pre-trained Depth-Anything-V2-Base model using our proposed method on the soccer dataset introduced in Section 2. We compare our approach with the

Figure 1. Qualitative comparisons of Depth Anything V2[8] fine-tuned on the Virtual KITTI dataset, our method without post-processing (denoted as *Ours (w/o smoothing)*), and our full method (denoted as *Ours*). All predicted depth maps are aligned to the ground truth using scale and shift before visualization.

Method	Abs Rel$\times 10^{-3}$	RMSE$\times 10^{-3}$	RMSE Log$\times 10^{-3}$	Sq Rel$\times 10^{-4}$	SILog
Depth Anything V2	65.353	43.562	78.099	36.606	7.791
Ours (w/o smoothing)	1.252	1.395	2.196	0.048	0.220
Ours	**1.250**	**1.394**	**2.195**	**0.048**	**0.219**

Table 1. Quantitative comparison of depth estimation results

original Depth-Anything-V2-Large fine-tuned on the Virtual KITTI dataset as the baseline.

Figure 1 shows qualitative comparisons, and Table 1 presents quantitative evaluations. Since the evaluation metrics are computed after aligning the predicted depth maps with the ground truth using scale and shift to match their distributions, the qualitative results are also visualized after applying the same alignment.

As discussed in Section 2, the baseline Depth Anything V2 struggles to estimate accurate depths around players, often producing blurry and unreliable depth values in those regions. In contrast, our method significantly sharpens and improves the depth estimation around players by applying player-specific supervision. The post-processing step also contributes to further improvements, as shown in the quantitative results in Table 1.

References

[1] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy, 2024. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1

[3] Arnaud Leduc, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-depth: a scalable dataset for monocular depth estimation in sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3280–3292, 2024. 1

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[6] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024.

[7] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 1

[8] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2024. 1, 2