Player-Aware Monocular Depth Estimation for Soccer Broadcast Frames

Jacek Maksymiuk
maksymiukjacek1@gmail.com

April 17, 2025

Abstract

This paper presents our submission to the SoccerNet Monocular Depth Estimation (MDE) Challenge 2025[1], which aims to estimate depth maps from broadcast soccer frames. MDE in sports is a key enabler for understanding the 3D game state from 2D footage. We demonstrate that with targeted fine-tuning and depth correction focused on player regions, our method outperforms the provided baseline by 35–57% across different metrics. The complete codebase is available at: https://github.com/JacekMaksymiuk/football-depth-estimation

1. Introduction

Monocular Depth Estimation (MDE) from soccer broadcast footage is a challenging problem due to the dynamic nature of the scenes and the lack of camera parameters. The goal of this work is to develop a depth estimation pipeline that accurately estimates depth from a single RGB frame, with special attention to depth accuracy for soccer players.

While existing MDE models generalize well across standard datasets, they are not optimized for the specific characteristics of soccer broadcasts. In this paper, we fine-tune the DepthAnything model [5] for this task and introduce a secondary model to correct depth estimation within player regions. Player masks are obtained via the YOLOv8 segmentation model [1], and the correction model uses a UNet [3] with attention mechanism [4] to improve local accuracy.

Although global metrics see minor improvements from this correction step, visual and local accuracy within player regions improves significantly, which is crucial for tasks relying on accurate game state reconstruction.

2. Dataset

Our training dataset consists of 3,864 synthetic RGB-depth pairs rendered from video games. Each frame has a corresponding ground truth depth map. We use 1,441 pairs for validation during training and 1,423 real-world broadcast frames from the SoccerNet-Depth test set [2] for final evaluation.

[1] https://www.soccer-net.org/challenges/2025

3. Methodology

The overall pipeline consists of the following steps:

1. Apply a fine-tuned DepthAnything model to the input RGB frame.
2. Use YOLOv8 segmentation to identify player regions.
3. Refine depth predictions within these regions using a UNet with attention.

The final depth maps are saved as 16-bit PNGs.

3.1. Fine-tuning DepthAnything

The DepthAnything model was fine-tuned for 24 epochs using a batch size of 4.

3.2. Player Segmentation using YOLOv8

We use the pre-trained YOLOv8x-seg.pt model provided by Ultralytics [1] for segmentation. Player masks are used to extract crops for correction.

3.3. Depth Correction Model

We introduce an enhanced UNet-based model to correct depth predictions within player regions. The model is trained to regress the residual between the fine-tuned DepthAnything output and the ground truth depth.

Our architecture builds upon the classic UNet with several improvements: residual convolutional blocks in the encoder, dense blocks in the bottleneck for better feature reuse, and attention mechanisms in the decoder for sharper spatial focus. Specifically, we use squeeze-and-excitation (SE) modules to recalibrate channel-wise features, and a gated attention mechanism to modulate skip connections based on decoder context. These additions improve the network's ability to refine fine-grained depth within player regions, where accurate reconstruction is most critical.

Model	Abs Rel $\times 10^{-3}$	RMSE $\times 10^{-3}$	RMSE Log $\times 10^{-3}$	Sq Rel $\times 10^{-4}$	SILog
ZoeDepth	46.545	31.085	55.874	18.020	5.576
DepthAnything	4.105	3.680	6.130	0.262	0.613
DepthAnything-ft-sn	2.584	2.401	4.167	0.125	0.417
ZoeDepth-ft-sn	2.429	2.343	4.002	0.121	0.400
DepthAnything-ft (ours)	1.055	1.852	3.135	0.094	0.313
DepthAnything-player-ft (ours)	**1.044**	**1.510**	**2.525**	**0.055**	**0.252**

Table 1. Evaluation scores for all models on the SoccerNet-Depth test set. We report five standard depth metrics with scaled units: Absolute Relative Error (Abs Rel, $\times 10^{-3}$), Root Mean Squared Error (RMSE, $\times 10^{-3}$), RMSE in log space (RMSE Log, $\times 10^{-3}$), Squared Relative Error (Sq Rel, $\times 10^{-4}$), and Scale-Invariant Log Error (SILog). Our fine-tuned models (last two rows) outperform all baselines, with the final player-aware model (DepthAnything-player-ft) achieving the best performance across all metrics.

3.3.1. Training Data

Training samples are rectangular crops of player regions from the training set. Each crop is paired with its corresponding depth residual. This forms the dataset for training the correction model.

3.3.2. Training Setup

Training details and hyperparameters are available in the source code: https://github.com/JacekMaksymiuk/football-depth-estimation

4. Results

Table 1 presents a comparison of all models. ZoeDepth and DepthAnything represent baseline models, while DepthAnything-ft-sn and ZoeDepth-ft-sn are the SoccerNet fine-tuned versions. Our models — DepthAnything-ft and DepthAnything-player-ft — show significant improvements. Qualitative results (Figure 1) and player-specific results (Figure 2) highlight the benefits of the correction model.

Our model improves over the SoccerNet baseline by over 35–57% across different metrics. Player-aware correction further improves accuracy, particularly in areas critical for downstream tasks like tracking and 3D pose estimation.

5. Conclusion

This work demonstrates that fine-tuning a general-purpose MDE model on soccer-specific data, combined with a dedicated player-aware correction model, leads to significant performance gains. The correction step, while locally focused, provides crucial accuracy improvements within player regions, which are not captured by global metrics alone.

References

[1] Glenn Jocher, Ayush Chaurasia, Ting Chen, Laughing, Abhiram V, Alex Stoken, Jirka Borovec, Yonghye Kwon, Jebastin Nadar, Aditya Kadu, et al. Ultralytics yolov8: Cutting-edge object detection models. https://github.com/ultralytics/ultralytics, 2023. Accessed: 2025-04-15. 1

[2] 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 *CVPRW*, 2024. 1

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 9351:234–241, 2015. 1

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[5] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 1

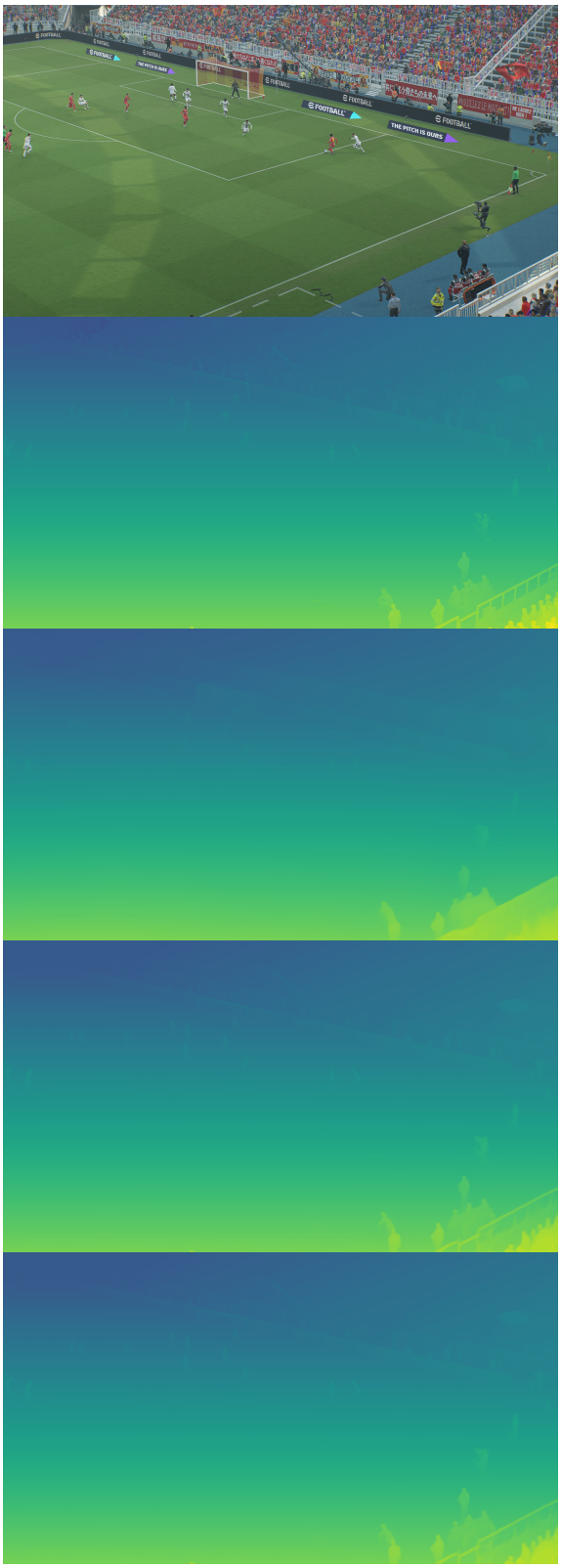

Figure 1. Qualitative comparison of depth maps generated by different models. From left to right: (1) original RGB input frame, (2) ground truth depth map, (3) prediction by the original DepthAnything model, (4) prediction by the fine-tuned DepthAnything model (DepthAnything-ft), and (5) prediction by our final model with player-aware correction (DepthAnything-player-ft).

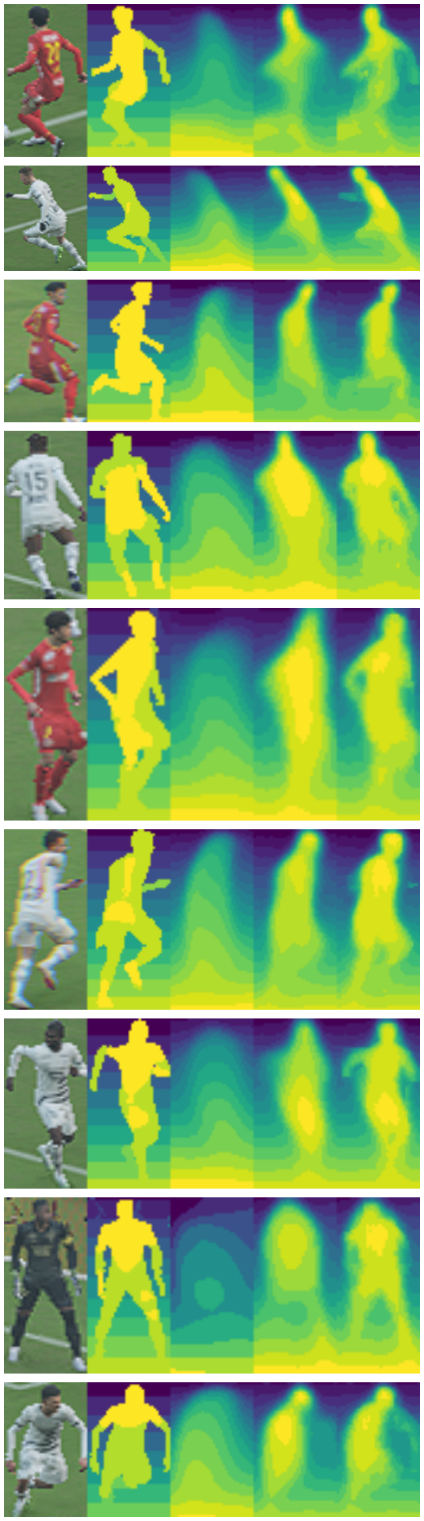

Figure 2. Player-focused qualitative comparison. Each row presents a zoomed-in player region, with five visualizations per row: (1) the original RGB crop, (2) the ground truth depth map, (3) output from the original DepthAnything model, (4) output from the fine-tuned DepthAnything model (DepthAnything-ft), and (5) prediction from our final model with player-aware correction (DepthAnything-player-ft). The improved depth accuracy and boundary sharpness within player regions demonstrate the effectiveness of our fine-tuning and correction stages.