HSLVision: A Multimodal Vision Dataset for RoboCup Humanoid Soccer

D. M. Xavier Catarrinho[1,2], G. de Jong[1,2], M. J. Meijer[1,2], H. Ruiter[1,2], and M. Honkoop[1,2]

[1] University of Amsterdam
[2] whIRLwind Amsterdam

Abstract. RoboCup 2026 marks the introduction of the Humanoid Soccer League (HSL), unifying the former Standard Platform League (SPL) and Humanoid League (HL). Existing data from these former leagues poorly covers the broader range of humanoid embodiments and viewpoints of the HSL. We present HSLVision, an RGB-D object detection dataset collected under HSL conditions, comprising 6,068 images and 65,309 annotations gathered across seven locations and events. Each frame is paired with a pixel-aligned monocular depth map and annotated for seven classes commonly used in autonomous play: Ball, Robot, Goalpost, Penalty mark, and T-, L-, and X-intersection. Alongside the data we define a reproducible benchmark with standardized splits and evaluation metrics, and provide eight baseline detectors drawn from the YOLO and DETR families. Baselines trained on HSLVision reach strong in-domain performance and are further evaluated on a held-out competition venue not seen during training, establishing a reference point for HSL perception research.

Keywords: Humanoid robot soccer · Dataset · Benchmark · Object detection · Monocular depth estimation · RoboCup HSL

1 Introduction

The Robot World Cup Initiative (RoboCup) is an international research competition aimed at advancing robotics, with the long-term goal of defeating the top human soccer team by 2050 [9]. Starting in 2026, the Standard Platform League (SPL), in which teams of identical NAO robots compete, and the Humanoid League (HL), which allows varied humanoid embodiments, are unified into the Humanoid Soccer League (HSL). The HSL is organized into three size-based divisions: Small (<110 cm), Middle (<125 cm), and Large (<190 cm) [18,17].

This unification breaks assumptions that existing perception pipelines rely on. SPL pipelines were tightly tuned to NAO-specific conditions: a consistent body silhouette, fixed camera height, and known field-of-view geometry. HL pipelines, while designed for varied embodiments, are largely based on older platforms. The HSL combines embodiment diversity with new field sizes, so perception has to be built against in-domain HSL data rather than borrowed from

either predecessor league. In this paper we introduce HSLVision, the reference object-detection dataset and benchmark for the HSL.

This paper makes two contributions:

- **Dataset.** HSLVision, the first public RGB-D detection dataset for the HSL: 6,068 images with paired depth maps and 65,309 annotations over seven classes, spanning multiple venues and robot embodiments.
- **Benchmark.** Standardized splits, evaluation protocol, and eight YOLO- and DETR-family baselines, reported in-domain and on a held-out competition venue.

2 Related work

In this section we elaborate on the existing datasets and vision pipelines within the RoboCup.

2.1 Soccer perception in RoboCup

Research on perception for robot soccer spans ball detection, robot detection, field localization, and higher-level scene understanding. Early SPL systems relied on hand-crafted color and shape heuristics, but the 2016 switch from the orange ball to a black-and-white patterned ball accelerated the adoption of CNN-based detectors [12,13,2]. Subsequent work has focused on adapting lightweight architectures to the embedded-compute constraints of robot soccer. Yao et al. release YOLO-LITE variants tuned for CPU-only inference on NAO-class hardware[21]. Robot detection has received comparatively less attention, but Poppinga and Laue propose JET-Net, a compact full-image detector that runs in real time on the NAO V5 and targets robot detection specifically [15].

2.2 Datasets and benchmarks in robot vision

Modern object detection was shaped by general-purpose benchmarks such as PASCAL VOC [4] and COCO [11], however neither captures the visual conditions of robot soccer. Within RoboCup, the SPL Object Detection Dataset V2 covers four SPL classes: Ball, Robot, Goalpost, and Penalty mark [21,22]. The closest precedent to our work is TORSO-21 [1], which combines real-world and Webots-simulated images from the HL with a smaller SPL subset. TORSO-21 demonstrates the value of multi-league benchmarking, but its images reflect older HL platforms.

3 HSLVision

This section explains HSLVision, the dataset for the HSL. It explains the contents of the dataset, preprocessing, annotation protocols, and it shows dataset statistics.

3.1 Design goals

HSLVision is designed around three goals. First, it should support the transition from SPL and HL to HSL conditions by providing annotated data that reflects the visual variability of the new league, including variation in robot embodiment, viewpoint, and scene structure across the Small, Middle, and Large HSL divisions [18]. Second, it should cover the full set of visual entities that HSL perception pipelines depend on: the ball and other robots drive ball-based play and collision avoidance, while goalposts and field marks drive self-localization. Accordingly, HSLVision annotates seven classes: Ball, Robot, Goalpost, Penalty mark, T-intersection, L-intersection, and X-intersection. Third, it should enable reproducible benchmarking by providing standardized splits, a consistent annotation policy, and clear evaluation protocols so that results are comparable across methods and teams.

3.2 Data collection

Images in HSLVision were collected across six locations: LAB42 (whIRLwind), German Open 2026 Cologne (whIRLwind), RCAP Beijing Masters 2025 (whIRLwind, HTWK), RoboCup 2025 Salvador (HTWK), RoboCup 2019 Sydney (BitBots), and RoboCup 2017 Nagoya (BitBots). In total, the dataset contains 6,068 images and 65,309 annotations. It includes a range of field layouts, goalpost designs, ball sizes, and robot platforms. The robots in the dataset include the Booster Robotics K1 and T1, as well as several non-standard KidSize humanoids. The data recorded in LAB42 contains images under various natural lighting conditions. The dataset includes data on the small, middle, and large field sizes, according to the HSL rules [19]. Table 1 and Table 2 summarizes the key properties of HSLVision.

In addition to this, we used 770 annotated images from the RCAP Abu Dhabi 2025 competition to evaluate the performance on unseen locations.

Recording was carried out using Booster K1s, Booster T1s, and a custom humanoid platform.

Table 1: Annotations per class.

Class name	**#Annotations**
Ball	3,368
Goalpost	7,084
Robot	14,996
L-intersection	17,918
Penalty mark	3,820
T-intersection	13,117
X-intersection	5,006
Total	**65,309**

Table 2: Dataset breakdown by source.

Dataset	**#Images**	**#Annotations**
LAB42	1,337	16,441
RoboCup 2025	1,506	19,289
RCAP Beijing 2025	2,313	23,045
GO 2026	764	5,473
RoboCup 2019	120	788
RoboCup 2017	28	273
Total	**6,068**	**65,309**

3.3 Data selection

As illustrated in Figure 1, raw footage was first subsampled at 3 fps. The resulting images were then filtered before annotation. Frames with excessive motion blur or occlusion, corrupted images, and low-quality images were removed. After initial filtering, we remove visually similar frames. Fixed-rate video contains many near-identical images, which waste annotation effort and bias training toward over-represented viewpoints. These redundant frames are discarded before labeling. Our approach is inspired by the TORSO-21 dataset [1], which removes similar frames based on distances in a learned latent space. We follow the same idea by embedding each image and discarding frames that are too close in embedding space, but use a different encoder and selection strategy. Instead of training a dataset-specific VAE [8], we use a pretrained DINOv3 ViT-S/16 encoder [20] and take the 384-dimensional pooler token as the image embedding. Embeddings are L2-normalized so that cosine similarity reflects distance. Given a similarity threshold τ, we iteratively remove the image with the largest number of above-threshold neighbors until no remaining pair exceeds τ, as illustrated in Figure 1 (bottom). This yields a diverse subset without depending on frame order. Using a pretrained encoder avoids per-dataset training and provides a semantically meaningful embedding, which is important for small scene-specific datasets.

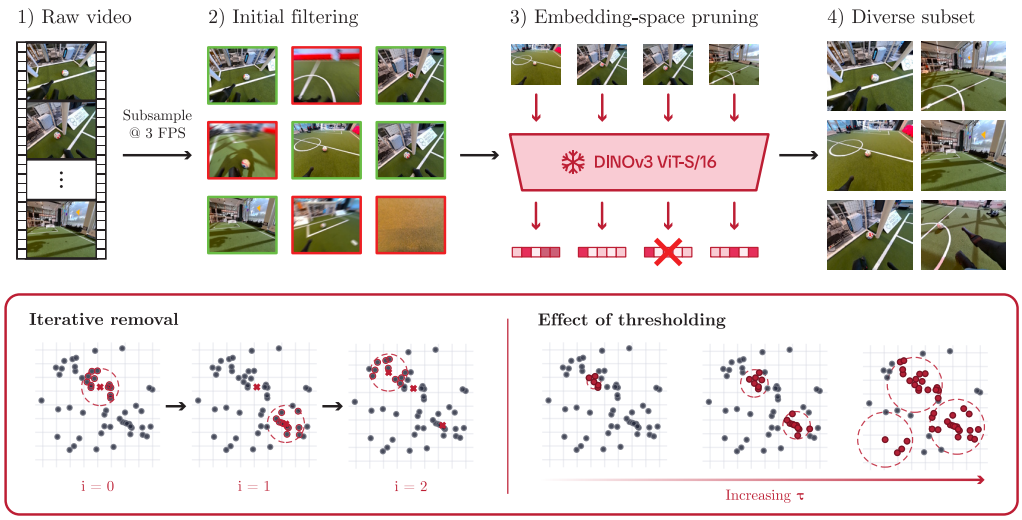

Fig. 1: Overview of the data selection pipeline. Bottom: illustration of the embedding-space pruning procedure, where points with many nearby neighbors are iteratively removed, and the effect of varying τ on the diversity of the selected subset.

3.4 Depth map

Some recording platforms, such as the Booster K1, provide onboard depth estimates using a stereo pipeline based on a D-Robotics stereo camera and the Hobot StereoNet[3] model. However, raw depth outputs are not available across all recording platforms, as not all of them include a stereo camera setup. To ensure uniform coverage, they are not included in the dataset.

Instead, every color frame is paired with a depth map estimated using DepthAnythingV3 [10], which predicts per-pixel metric depth from a single RGB image. The resulting depth maps are aligned with the RGB images and share the same pixel coordinates, so annotations can be applied directly to both modalities without additional transformation. This is illustrated in Figure 2. Qualitatively, DepthAnythingV3 produces noticeably smoother and more spatially consistent depth maps than the Hobot StereoNet baseline. Object silhouettes such as the ball in the field scene and the goalposts are preserved with sharp, well-localized boundaries. Despite relying only on a single RGB input, the metric depth predicted by DepthAnythingV3 appears to be on par with, if not more accurate than, the stereo baseline.

Fig. 2: Comparison of RGB input, Hobot StereoNet, and DepthAnythingV3 predictions across two scenes.

3.5 Annotation protocol

Annotations cover seven classes, grouped into *objects* (Ball, Robot, Goalpost) and *field marks* (Penalty mark, T-intersection, L-intersection, X-intersection).

All instances are annotated with axis-aligned bounding boxes. For field marks, the bounding box tightly encloses the mark or line intersection, with the intersection point at the center of the box. Field marks are labeled only when their type can be determined from the image alone. Objects are annotated when at least 25% of the object is visible. Figure 3 shows an example of a completely

[3] `https://github.com/D-Robotics/hobot_stereonet` based on StereoNet [7]

annotated image. Annotations were created using a custom pipeline built on the CVAT [3] annotation tool.

Fig. 3: Example of fully annotated image from dataset, taken at the World Humanoid Robot Games 2025 in Beijing (RCAP Beijing Masters 2025).

4 Benchmark

This section explains the first benchmarks applied to HSLVision, in the form of four detector families at two size variants. We provide a standardized test set that can be used for comparisons to future approaches.

4.1 Metrics

Detection performance is reported using standard COCO-style metrics [11]: mean average precision (AP) at IoU threshold 0.5 (AP_{50}), at 0.75 (AP_{75}), and averaged over IoU thresholds from 0.5 to 0.95 in steps of 0.05 (AP). Object-tier and field mark-tier AP are reported both jointly and separately, since the AP of the field marks is more sensitive to localization noise. We also report the AP across three different object scales (AP_S, AP_M, and AP_L), following the COCO evaluation protocol [11]. This breakdown is particularly relevant for field marks, which predominantly fall into the small-object regime (AP_S).

4.2 Baseline models

We evaluate four detector families at two size variants each, giving eight baselines that span a range of speed–accuracy trade-offs relevant to robotic deployment. The YOLO family is represented by YOLOv11 [5] and YOLOv26 [6], both convolutional real-time detectors. The DETR family contributes two transformer-based real-time detectors: RF-DETR [16], and D-FINE [14]. For each family we train the nano (N) and small (S) variants. All baselines are finetuned from their published pretrained weights on the HSLVision training split. Hyperparameters follow each architecture's published defaults with a fixed random seed.

4.3 Results

Table 3 shows the results of the benchmarks on the test split. For both the nano and the small variant, D-FINE outperforms all other models. The performance on large bounding boxes is similar for all models. However, the difference in performance for small and medium bounding boxes is substantial. This pattern is also apparent when comparing the different models on field mark performance, where D-FINE is considerably better than the other models.

Table 3: Overall detection performance on the test set. AP_{obj} averages Ball, Robot, and Goalpost; AP_{mark} averages the L-, T-, X-intersections and Penalty marks. Best value is denoted in **bold**.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP_{obj}	AP_{mark}
nano variants								
YOLO11-N	0.724	0.952	0.794	0.647	0.785	0.809	0.795	0.671
YOLO26-N	0.726	0.954	0.800	0.656	0.778	0.773	0.795	0.674
RF-DETR-N	0.736	0.963	0.811	0.662	0.785	**0.830**	0.798	0.690
D-FINE-N	**0.768**	**0.966**	**0.830**	**0.731**	**0.840**	0.803	**0.819**	**0.729**
small variants								
YOLO11-S	0.752	0.963	0.818	0.686	0.801	**0.827**	0.821	0.700
YOLO26-S	0.758	0.966	0.826	0.703	0.796	0.807	0.820	0.712
RF-DETR-S	0.762	0.970	0.827	0.704	0.803	0.825	0.822	0.716
D-FINE-S	**0.790**	**0.971**	**0.847**	**0.763**	**0.845**	0.803	**0.845**	**0.748**

Table 4 summarizes the performance of the D-FINE models across different distances. Both the nano and the small variant perform best at mid-range distances from one to four meters. For farther away boxes, the performance drops more for objects, while the AP of the field marks stays more stable. A possible explanation for this could be the larger amount of far away field marks in the

dataset. For objects within one meter of the camera, the performance is considerably worse than for objects slightly farther away. However, the number of ground truth objects that are close to the camera is very small, suggesting that the drop in performance is because of a lack of representation of these objects in the dataset. This, in turn, can be explained by the lack of situations in which detection is required at ranges between 0 and 1m in a match situation.

Table 4: Detection performance on the test set, specified for different distances in meters. AP_{obj} averages Ball, Robot, and Goalpost; AP_{mark} averages the L-, T-, X-intersections and Penalty marks. Best value is denoted in **bold**.

	#GT objects	AP	AP_{50}	AP_{75}	AP_{obj}	AP_{mark}
D-FINE-N						
0-1 m	87	0.612	0.826	0.606	0.476	0.794
1-2 m	588	**0.837**	0.964	**0.917**	**0.820**	**0.860**
2-4 m	2,188	0.832	**0.967**	0.907	0.816	0.852
4-8 m	4,173	0.744	0.965	0.805	0.700	0.801
8+ m	2,435	0.657	0.926	0.715	0.622	0.704
D-FINE-S						
0-1 m	87	0.614	0.817	0.579	0.488	0.783
1-2 m	588	**0.852**	0.959	**0.924**	**0.834**	0.876
2-4 m	2,188	0.847	0.972	0.913	0.823	**0.879**
4-8 m	4,173	0.767	**0.973**	0.831	0.720	0.830
8+ m	2,435	0.699	0.950	0.748	0.656	0.756

4.4 Evaluation on out-of-distribution data

A key aspect of the RoboCup tournament is that, every year, it is hosted in a different location. Because of this, it is essential that vision models are robust and perform well on out-of-distribution data. To benchmark this, all models were evaluated on a separate dataset recorded at RCAP Abu Dhabi 2025, which contains 770 images. This dataset provides unseen fields and backgrounds.

Table 5 shows that there is a large performance drop when evaluating the trained models on data from an unseen location. Among the nano variants, RF-DETR outperforms all other models, suggesting that while D-FINE scores better on seen locations, it is not able to adapt as well to unseen locations. The small variants of RF-DETR and D-FINE perform similar, whilst D-FINE is considerably better at small objects but worse on large objects. For both cases, RF-DETR is more accurate in object detection while D-FINE detects field marks more accurately.

Table 5: Overall detection performance on a 770-image dataset at a location not present in the training set. AP_{obj} averages Ball, Robot, and Goalpost; AP_{mark} averages the L-, T-, X-intersections and Penalty marks. Best value denoted in **bold**.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP_{obj}	AP_{mark}
nano variants								
YOLO11-N	0.446	0.729	0.437	0.267	0.485	0.437	0.573	0.351
YOLO26-N	0.440	0.717	0.449	0.273	0.480	0.393	0.551	0.356
RF-DETR-N	**0.462**	**0.740**	**0.459**	0.228	**0.497**	**0.572**	**0.591**	0.365
D-FINE-N	0.424	0.692	0.424	**0.346**	0.484	0.389	0.501	**0.367**
small variants								
YOLO11-S	0.459	0.730	0.468	0.287	0.503	0.453	0.580	0.369
YOLO26-S	0.462	0.737	0.472	0.309	0.496	0.403	0.589	0.366
RF-DETR-S	**0.479**	**0.760**	0.485	0.280	0.509	**0.580**	**0.621**	0.372
D-FINE-S	0.478	0.735	**0.489**	**0.404**	**0.554**	0.400	0.560	**0.416**

5 Release and reproducibility

HSLVision is publicly available on Hugging Face[4] under the CC BY 4.0 license. The project website[5] contains code for training, evaluation, dataset tooling, and depth-map generation.

6 Conclusion

We presented HSLVision, an RGB-D object detection dataset and benchmark for the RoboCup Humanoid Soccer League. The dataset contains 6,068 images and 65,309 annotations across seven classes, collected from multiple locations and robot platforms. It provides standardized splits, evaluation metrics, and baseline results for HSL perception.

The baseline results show strong in-domain performance, but also reveal clear challenges in generalizing to unseen venues, especially for field marks. By releasing the dataset, depth maps, annotations, and evaluation code, HSLVision provides shared infrastructure for developing and comparing HSL vision systems.

Acknowledgments. We thank all members of team whIRLwind Amsterdam for their annotation efforts, and team HTWK for providing some of the raw recordings used in this work.

[4] `https://huggingface.co/datasets/whirlwind-ams/hslvision`
[5] `https://intelligentroboticslab.github.io/hslvision/`

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bestmann, M., Engelke, T., Fiedler, N., Güldenstein, J., Gutsche, J., Hagge, J., Vahl, F.: TORSO-21 dataset: Typical objects in RoboCup soccer 2021. In: RoboCup 2021: Robot World Cup XXIV. Lecture Notes in Artificial Intelligence, vol. 13132. Springer (2022)
2. Cruz, N., Lobos-Tsunekawa, K., Ruiz-del Solar, J.: Using convolutional neural networks in robots with limited computational resources: Detecting NAO robots while playing soccer. In: RoboCup 2017: Robot World Cup XXI. Lecture Notes in Artificial Intelligence, vol. 11175, pp. 19–30. Springer (2018). https://doi.org/10.1007/978-3-030-00308-1_2
3. CVAT.ai Corporation: Computer Vision Annotation Tool (CVAT) (Nov 2023), https://github.com/cvat-ai/cvat
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4
5. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), https://github.com/ultralytics/ultralytics
6. Jocher, G., Qiu, J.: Ultralytics yolo26 (2026), https://github.com/ultralytics/ultralytics
7. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: Proceedings of the European conference on computer vision (ECCV). pp. 573–590 (2018)
8. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014)
9. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E.: Robocup: the robot world cup initiative. Proceedings of the International Conference on Autonomous Agents (04 1998). https://doi.org/10.1145/267658.267738
10. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=yirunib8l8
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48
12. Menashe, J., Kelle, J., Genter, K., Hanna, J., Liebman, E., Narvekar, S., Zhang, R., Stone, P.: Fast and precise black and white ball detection for RoboCup soccer. In: RoboCup 2017: Robot World Cup XXI. Lecture Notes in Artificial Intelligence, vol. 11175, pp. 45–58. Springer (2018). https://doi.org/10.1007/978-3-030-00308-1_4
13. O'Keeffe, S., Villing, R.C.: A benchmark data set and evaluation of deep learning architectures for ball detection in the RoboCup SPL. In: RoboCup 2017: Robot World Cup XXI. Lecture Notes in Artificial Intelligence, vol. 11175, pp. 398–409. Springer (2018). https://doi.org/10.1007/978-3-030-00308-1_33

14. Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., Wu, F.: D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=MFZjrTFE7h
15. Poppinga, B., Laue, T.: Jet-net: real-time object detection for mobile robots. In: Robot World Cup, pp. 227–240. Springer (2019)
16. Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: RF-DETR: Neural architecture search for real-time detection transformers. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=qHm5GePxTh
17. RoboCup Federation: Robocupsoccer standard platform league (spl). https://spl.robocup.org/ (2025), accessed: 2026-04-24
18. RoboCup Federation: Humanoid Soccer League: Call for Participation. https://www.robocup.org/news/187 (2026), published January 8, 2026; accessed April 19, 2026
19. RoboCup Humanoid Soccer League Technical Committee: RoboCup Humanoid Soccer League Rules Repository. https://github.com/RoboCup-HumanoidSoccerLeague/HSL-Rules/releases (2026), official releases page for the 2026 HSL rules; accessed April 19, 2026
20. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
21. Yao, Z., Douglas, W., O'Keeffe, S., Villing, R.: Faster YOLO-LITE: Faster object detection on robot and edge devices. In: RoboCup 2021: Robot World Cup XXIV. Lecture Notes in Artificial Intelligence, vol. 13132, pp. 226–237. Springer (2022). https://doi.org/10.1007/978-3-030-98682-7_19
22. Yao, Z., Douglas, W., O'Keeffe, S., Villing, R.: SPL Object Detection Dataset V2. https://roboeireann.maynoothuniversity.ie/research/SPLObjDetectDatasetV2.zip (2022), released with [21]