SoccerNet-Depth: a Scalable Dataset for Monocular Depth Estimation in Sports Videos

Arnaud Leduc[1] Anthony Cioppa[1,2] Silvio Giancola[2] Bernard Ghanem[2] Marc Van Droogenbroeck[1]
[1] University of Liège [2] KAUST
arnaud.leduc@student.uliege.be, anthony.cioppa@uliege.be, silvio.giancola@kaust.edu.sa
Bernard.Ghanem@kaust.edu.sa, M.VanDroogenbroeck@uliege.be

Abstract

Monocular Depth Estimation (MDE) is fundamental in sports video understanding, enhancing augmented graphics, scene understanding, and game state reconstruction. Despite remarkable progress in autonomous driving and indoor scene understanding, there is currently a lack of MDE datasets tailored for sports. Furthermore, most existing datasets only focus on single images, disregarding the temporal aspect. In this work, we introduce the first video dataset for MDE in sports, SoccerNet-Depth*, focusing on football and basketball videos. In particular, we leverage the graphic engine from video games to automatically extract video sequences and their associated depth maps, making our dataset easily scalable. Furthermore, we benchmark and fine-tune several state-of-the-art MDE methods on our dataset. Our analysis shows that MDE in sports is far from being solved, making our dataset a perfect playground for future research. Dataset and codes: https://github.com/SoccerNet/sn-depth.*

1. Introduction

Deep learning brought significant advancements in the field of computer vision, allowing a comprehensive analysis of images and videos. A critical focus area is *depth estimation*, whose objective is to determine the real-world distance of every object in a scene to the camera. This fundamental aspect allows for a more in-depth understanding of the spatial relationship between the objects and the environment. In practice, the distance can be estimated in *relative depth*, which captures the order and spatial relationships among objects without explicit distance measures, or in *metric depth*, which quantifies the exact distances from the camera to the objects in real-world units. Estimating depth maps of images or videos can be achieved through several approaches. Geometry-based methods typically leverage motion or multiple points of views to determine the depth [33]. Sensor-based approaches take advantage of Time-of-Flight (ToF) or Lidar technology [31]. More recently, deep learning techniques enabled monocular depth estimation, *i.e.*, injecting domain knowledge to estimate depth from a single point of view.

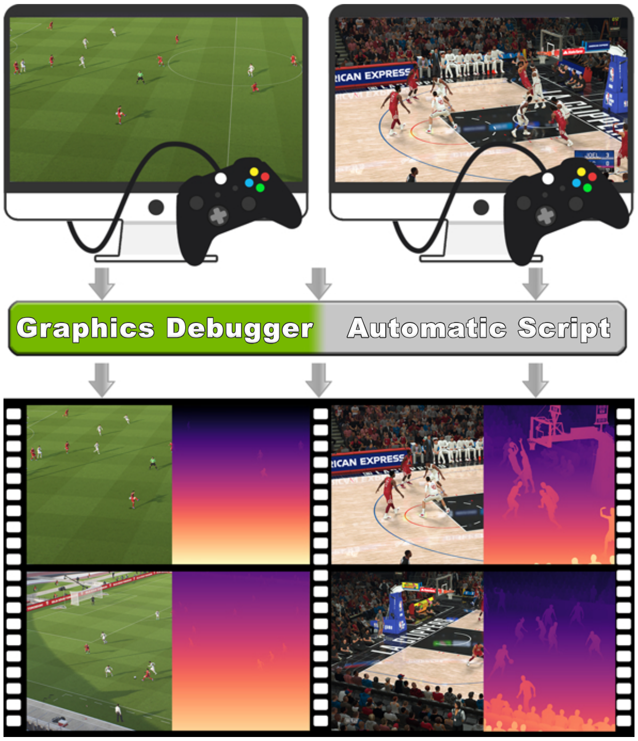

Figure 1. **SoccerNet-Depth.** We introduce a novel scalable dataset for Monocular Depth Estimation in sports videos. Our synthetic dataset is generated from video games, simulating football and basketball games. We leverage graphics debuggers and automatic scripts to extract video sequences along with their depth maps.

In the world of sports, where analytics play a crucial role in enhancing performance, the information brought by Monocular Depth Estimation (MDE) offers interesting possibilities. First, the monocular video modality is available for all recorded games, providing a cheaper option for depth estimation compared to expensive sensors or multi-view setups. Second, by providing a three-dimensional perspective from two-dimensional video captures, MDE significantly improves scene understanding, such as allowing player and ball tracking in 3D. Furthermore, depth maps can be used to enhance the broadcast with the integration of Augmented Reality (AR) content between the players, or by adding depth-of-field blur to give a cinematic look to the footage. However, state-of-the-art deep learning methods are typically trained on annotated depth data, and yet, those data remain scarce in sports. Indeed, gathering depth data is challenging for real-world sports matches, as depth information is usually captured through ToF sensors such as LiDAR and Kinect devices. While these methods are accurate in most scenarios, they are often obstructed by their high-cost and low data acquisition rates, in addition to requiring sophisticated setup and calibration processes. Hence, huge playing fields and high-speed game dynamics limit their accessibility and widespread adoption in the sports community.

As an alternative solution, we focus on synthetic data generation, which allows extracting various information automatically. Nowadays, video games and sports simulations have reached high-level realism, with improved graphics, realistic ball dynamics, and advanced AI-driven player movements. Leveraging the graphics engine pipeline allows extracting numerous videos and their computed depth maps at a low cost. In this work, we introduce and benchmark a novel dataset, *SoccerNet-Depth*, composed of synthetic video sequences alongside their corresponding depth maps. The data are extracted from two popular video games: *NBA2K22* and *EFootball*, as illustrated in Figure 1. Our approach is scalable, both in size and number of sports, and complements previous efforts in MDE by providing a meaningful dataset to train current methods on sports videos.

Contributions. We summarize our contributions as follows. **(i)** We propose *SoccerNet-Depth*, the largest publicly available dataset for monocular depth estimation on team sports videos, with 12,398 pairs of synthetic frames and depth maps automatically extracted from football and basketball video games. **(ii)** We benchmark and fine-tune several state-of-the-art monocular depth estimation methods on our new *SoccerNet-Depth* dataset, showcasing the remaining challenges for future research.

2. Related Work

2.1. Monocular Depth Estimation

Methods. Significant progress have been noted in the field of Monocular Depth Estimation (MDE). The early foundational work by Saxena *et al*. [77] employed Markov Random Field (MRF) to predict depth from image features. The field further evolved significantly with the adoption of deep learning techniques, particularly Convolutional Neural Networks (CNNs) [19, 26, 37, 53, 56]. A significant improvement was then brought by BTS [47], which enhanced existing models by adding new local planar guidance layers in the network, setting unprecedented records at the time.

More recently, methods based on Transformer [91] were introduced in MDE, leveraging non-local attention-based aggregation in contrast with CNNs. The DepthFormer model [50] exemplifies this trend, combining the strengths of both Transformer and CNN models. Another significant breakthrough was achieved with Ranftl *et al*.'s DPT model [70], which employs vision transformers, diverging from conventional convolutional networks. This model, drawing inspiration from the ViT model by Dosovitskiy *et al*. [18], sets a new state of the art in the field. Afterward, Bhat *et al*. introduced Adabins [21] achieving groundbreaking performance by coupling a standard encoder-decoder block to a new transformer-based architecture that splits the depth space into bins. Innovative methods later built on top of Adabins, such as BinsFormer [52] and LocalBins [4], pushed the performance even further. Alternatively, depth completion techniques [65,98,102] utilize sparse depth data to generate detailed dense depth maps.

Recently, several works [56, 97, 100] investigated temporal consistency for depth estimation in video sequences. Luo *et al*. [56] developed a novel system for calculating depth from monocular videos, ensuring both temporal consistency and geometric accuracy. NeWCRFs [100] leverages fully-connected Conditional Random Fields (CRFs) and multi-head attention to predict sequences by modeling the dependencies between their constitutive elements, improving the contextual accuracy of predictions. MAMo [97] also introduced temporal consistency to perform video depth estimation. To do so, the work relies on other models, such as NeWCRFs [100] or PixelFormer [1]. The latter is based on a skip attention method that facilitates efficient information flow across different layers of a neural network by allowing layers to skip connections, enhancing the learning of both high-level and low-level features.

MiDaS by Ranftl *et al*. [71] enabled training on multiple datasets simultaneously [45, 51, 54, 93] and achieved excellent overall performance on unseen datasets. MiDaS [71] serves as a foundational step in ZoeDepth [5], a prominent monocular depth estimation method, that integrates relative and metric depth estimation to enhance the performance. ZoeDepth [5] was a turning point for highly performing methods such as PatchFusion [49] and Depth-Anything [96]. Finally, diffusion models [43, 46, 103] show good performances on popular datasets. These approaches might significantly improve the field in the future. In this work, we benchmark and fine-tune several state-of-the-art MDE methods on our new SoccerNet-Depth dataset.

Datasets. Typically, deep learning algorithms require large

Table 1. **Monocular depth estimation datasets comparison.** SoccerNet-Depth is the largest publicly available dataset for monocular depth estimation on team sports videos, with 12.4k pairs of synthetic frames and depth maps from football and basketball video games.

Dataset	**Data Source**	**Data Type**	**Scenario**	**Images**	**Resolution**	**Depth**	**Video**	**Public**
DIML [45]	Kinect-V2/Zed	Real	Indoor/Outdoor	2M	C:1920×1080 D:512×424	Metric	✓	✓
NYUv2 [82]	Kinect-V1	Real	Indoor	1,449	640×480	Metric	✓	✓
KITTI [28]	LiDAR	Real	Driving	93k	1024×320	Metric	✓	✓
Diode [90]	Laser Scanner	Real	Indoor/Outdoor	25.5k	1024×768	Metric	✓	✓
MegaDepth [51]	SfM and MVS	Internet Photos	Outdoor	130k	Various	Euclidean + Ordinal	X	✓
Mid-Air [23]	Unreal Engine/Airsim	Synthetic	Drone	420k	1800×1800	Metric	✓	✓
UnrealStereo4K [87]	Unreal Engine	Synthetic	Outdoor/Indoor	8k	3840×2160	Stereo	X	✓
MVS-SYNTH [37]	Game	Synthetic	Outdoor	12k	1920×1080	Metric	✓	✓
MADS [101]	Stereo	Real	Individual Sports	5,855	1024×768	Metric	✓	✓
Soccer on Your Tabletop [72]	Game	Synthetic	Football	12k	256×256	Metric	✓	X
SoccerNet-Depth	Game	Synthetic	Team Sports	12.4k	1920×1080	Relative	✓	✓

datasets, consisting of diverse scenes with precise ground truth labels, to train on. Monocular depth estimation is not an exception and the field can rely on different kind of dedicated datasets. The most popular publicly available depth datasets, NYUv2 [82], made of indoor scenes, and KITTI [28], captured in driving scenarios, consist of real-world color images associated with ground-truth depth maps. Despite their drawbacks, such as low resolution (NYUv2) and sparsity (KITTI), they remain popular benchmarks for any new MDE method. The DIODE dataset [90] is another great resource that was collected using a professional-grade LiDAR scanner to capture accurate depth measurements of diverse indoor and outdoor real scenes. Additionally, the SUN RGB-D [85] and the DIML Indoor [45] datasets contribute with real RGB-D images of indoor scenes and are completed by high resolutions data from the Middlebury 2014 [78] dataset. Interestingly, MADS [101] focuses on human pose tracking in sports, offering stereo-based depth images of various sports actions. Yet the actions are performed by a single athlete and captured from a single viewpoint in a controlled environment. Finally, obtaining reliable depth data is expensive and requires efficient technological tools. To battle those downsides, Li *et al*. [51] proposed a novel approach to build their depth estimation dataset, MegaDepth, using Internet photo collections as a data source. In this work, we propose a first dataset for monocular depth estimation in team sports videos, leveraging synthetic data, as highlighted in Table 1.

Synthetic datasets. Recently, researchers used video games to generate realistic images with depth data. These synthetic data are used to train models in simulation before transferring to real-world data [63, 68, 94]. In MVS-Synth [37], the authors generated diverse urban scenes from the *GTA5* video game. Particularly, they extracted 120 video sequences, each containing 100 color frames and their corresponding ground-truth disparity maps, as well as the camera parameters. Some follow-up works [69, 73] also used GTA5 as baseline to extract data for either depth estimation or semantic segmentation. Following a similar idea, Fonder *et al*. [22, 23] generated the Mid-Air dataset, a synthetic collection of low-altitude drone flight data captured in unstructured environments, created using an UAV simulator. Recently, Unreal Engine has enabled many research teams to easily acquire valuable data such as UnrealStereo4K [87]. Our dataset follows this current trend, leveraging sports video games as powerful simulators.

2.2. Sports video understanding.

Methods. Sports video understanding has been a prominent research topic in the past decade [62, 64, 86]. Recent developments enable the delivery of accurate, real-time data and insights into player performance [7, 80, 89], tactics [2], and game events [13, 36, 83], elevating coaching strategies and contributing to an improved viewer experience [60, 76]. This expansion encompasses areas like action spotting [11, 16, 17, 20, 25, 29, 41, 44, 75, 81, 95], segmentation and tracking of players or the ball [12, 24, 38, 58, 99], and the creation of video highlights or summaries [10, 27].

Datasets. The field has experienced significant growth, characterized by a diverse array of datasets [39, 40, 79, 88, 101] and tasks [34, 35, 55, 59, 61, 84, 92]. The SoccerNet datasets and challenges [15, 30] have contributed to multiple video understanding tasks in sports such as, action spotting [8, 29], replay grounding [16], camera calibration [57] and player re-identification [9], multiple player tracking [14], multi-view video recognition [34, 35], and dense video captioning [61]. This work extends the SoccerNet dataset by adding depth data for football and basketball.

Synthetic sport datasets. Isolated efforts have been made to link sport analysis and video games. Sheng *et al*. [48] delved into depth estimation in a football game, *FIFA Football World*, building a dataset of 6.5k pairs of RGB images and depth maps of football scenes extracted from the game. Similarly, Zhu *et al*. [104] offered insights in obtaining images and using them to perform player reconstruction from a basketball video game. The dataset was collected playing the NBA2K19 game and intercepting calls between the game engine and the graphics card using RenderDoc [42].

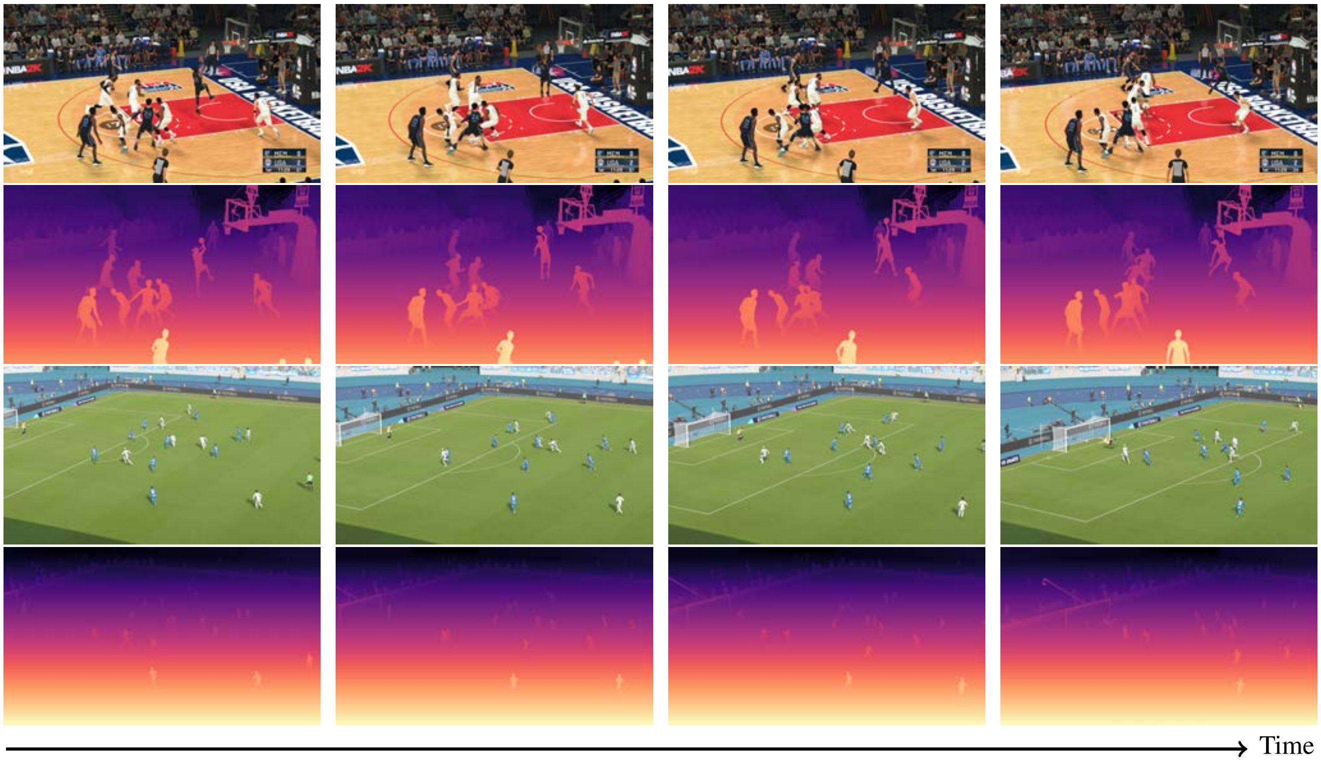

Figure 2. **Example of RGB frames and depth maps sequences in SoccerNet-Depth.** Top to bottom: **(1)** Basketball sequence showing a dunk with all players, referees and spectators in the camera field of view. **(2)** Associated ground-truth depth maps. **(3)** Football sequence including a shot and a save with multiple players appearing in the camera field of view. **(4)** Associated ground-truth depth maps.

Lastly, Soccer on your tabletop by Rematas *et al*. [72] explored methods to estimate the depth map of each player, using a CNN that is trained on 3D player data extracted from another football video game, FIFA. In our work, we also leverage video games to extract data and share our dataset with the research community.

3. SoccerNet-Depth

Our *SoccerNet-Depth* dataset consists of 74 synthetic video sequences generated from two widely renowned video games: *Efootball* and *NBA2k22*, respectively simulating football and basketball matches. Particularly, we extract two distinct types of sequences: synthetic RGB frames and associated depth maps. Each frame and 16-bits depth map is rendered at 1080p resolution (Full-HD), ensuring high-fidelity visual data. Examples of video sequences and their corresponding depth maps can be visualized in Figure 2.

Data collection. We played a total of 70 games with automatically piloted players at top skill level for both sports. To extract the frames and depth maps, we leveraged the deferred shading principle. This advanced rendering technique works by decoupling the shading process from geometry processing. Initially, scene geometry is rendered into multiple buffers. Subsequent stages involve applying lighting and shading effects. Therefore, by finding the appropriate buffer, depth information can be retrieved. RenderDoc [42] has been the prevalent tool to extract depth maps in various synthetic data research [48,72,73,104]. However, the author restricted its usage, even for research purposes. As an alternative, we chose NVIDIA Nsight [66] since our system's configuration incorporates an NVIDIA graphics card. During a typical gameplay, NBA2K runs at 115 frames per second (fps), whereas Efootball runs at 60 fps. Since our objective is to mirror real-world sequences, we wanted to simulate a 30 fps video output. However, Efootball and NBA2K limit frame accessibility through NVIDIA Nsight, restricting the frame rate capture capability to an upper limit of respectively, 1.2 and 6 fps.

Automatic data extraction. Manually extracting a single frame and its corresponding depth buffer takes on average 1.5 minutes for a person, which makes the manual approach intractable. To overcome this challenge, we developed a publicly available python script using libraries such as pyautogui, pydirectinput, and imagesearch [32] to automate the data extraction process. The script is programmed to interact seamlessly with NVIDIA Nsight [66], the Windows operating system and the video game environment. It employs imagesearch [32] to analyze the screen and iden-

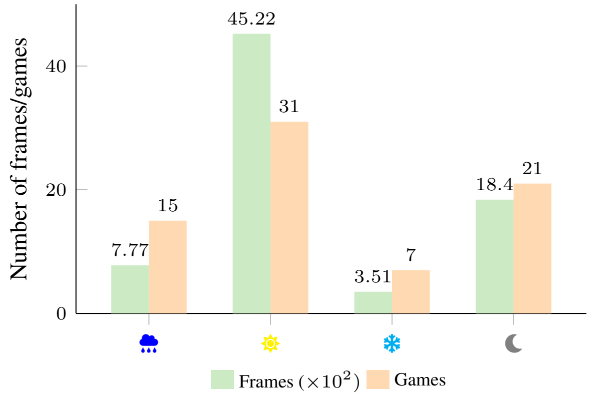

Figure 3. **Distribution of football frames and games per weather condition**. A frame can belong to more than one class.

tify precise click locations based on predefined images. As a result, we reduce the time per frame by approximately 30 seconds and automate the extraction, allowing for continuous operation, including overnight execution, effectively increasing the number of data extracted per day.

Data statistics. The 70 games, 62 of football and 8 of basketball, are split into train, test, and validation sets, publicly available, and a fourth challenge set, currently kept private for a future challenge, to prevent overfitting. The three first sets encompass a total of 12,398 frames, split following a 60/20/20 distribution with each game only appearing in one set. For football, there are 7,073 football frames, 4,071 for training, 1,423 for testing, and 1,579 for the validation set. For basketball, we provide a total of 5,325 basketball frames, 3,270 for training, 1,064 for testing, and 991 for validation. Extracting information from video games also brings the advantage of being able to control the external conditions of a match such as different weather (especially in football), including snow, rain, or sun, and either a day or night match. Figure 3 depicts the distribution of the various conditions in SoccerNet-Depth for the football dataset. Basketball being an indoor sport, the conditions remain the same across games, making it an easier dataset and benchmark. During the splitting of the football data, we ensured that each set has a comprehensive coverage of the different weather conditions and times of day.

Data format. The dataset is organized into one folder per sport, *i.e*., Efootball and NBA2K data. Within each section, the dataset is compartmentalized into distinct game folders, which in turn contains a series of video sequence folders. Each video sequence folder contains four subfolders named: *color*, *depth*, *depth_r* and *depth_buffer*. The *color* and *depth* folders contain PNG files, stored in 8-bits, with the depth scaled for visualization purposes. Conversely, the *depth_buffer* folders contains the raw CSV files with 16-bits depth information. This buffer keeps the temporal consistency across frames of a video sequence. We also provide the same information in a 16-bit PNG format in *depth_r*. In a video folder, files are systematically named according to their order of appearance in the clip, *i.e*., *[x].[png/csv]* for each frame *[x]*. At last, one .json file accompanies each game, providing metadata with contextual details. For football games, the metadata includes information about the weather, the shirt color of both teams, and the time of the day. For basketball games, which are played indoors, the metadata covers the color of the shirts and the floor. Additionally, for both sports, the metadata file provides details about the number of frames in the clip, the resolution of the frames, and the frame rate.

Novelty. *SoccerNet-Depth* is unique by its domain of application and its scalability. Table 1 shows that SoccerNet-Depth is the largest public dataset to provide depth estimation from team sports videos, with 12.4k frames. Moreover, unlike other MDE sport datasets [72,101], SoccerNet-Depth contains scene-centric data from in-match scenarios. The video sequences make it valuable for temporally consistent depth estimation, while the synthetic nature and automated extraction process makes the dataset scalable and the methodology transferable to other sports video game.

4. Benchmarks

Tasks. Monocular depth estimation (MDE) aims at predicting the depth, *i.e*., a notion of distance separating the objects of a scene to a camera, for each pixel of an image taken by this camera. A depth map is then defined as an image containing the depth information per pixel. In the literature, the depth map can contain two types of values, either metric depth values or relative depth values. The former is expressed in real-world units, while the latter is expressed in relative scale and is thus invariant to scaling operations. In our work, since video games depth data are only provided in relative scale, we predict a relative depth value for each pixel of the different frames of the video sequences.

Metrics. To evaluate the performance of the different models on our datasets, we consider the four metrics introduced by Eigen *et al*. [19]: the absolute relative error (Abs Rel), the squared relative error (Sq Rel), the root-mean-square-error (RMSE), and the root-mean-square error on the logarithm (RMSE log). Additionally, we use a scale invariant (SILog) metric [19] that measures the relationships between points in the scene without considering any absolute global scale. For valid comparisons, we apply a mask to exclude scoreboard pixels present in color images, as they do not appear in the ground truths. It is worth noting that since our dataset provides relative depth ground truths, the distributions of the predictions of the methods have to be aligned with the ground-truths distribution. To do so, we follow the scale and shift operation based on a least-square

Table 2. **Benchmark of state-of-the-art methods.** Evaluation of the state-of-the-art methods on the football and basketball test set. We test 5 methods in inference mode and fine-tune (-ft) two of them, ZoeDepth and DepthAnything, on our football or basketball data. For the two sports, the best value is highlighted in **bold** whereas the second best is written in *italics*.

Sport	Models	Abs Rel$\times 10^{-3}$	RMSE$\times 10^{-3}$	RMSE Log$\times 10^{-3}$	Sq Rel$\times 10^{-4}$	SILog
Football	PatchFusion [49]	69.464	46.663	82.451	40.642	8.224
	Marigold [43]	55.124	36.510	65.833	25.538	6.568
	ZoeDepth [5]	46.545	31.085	55.874	18.020	5.576
	MiDaS [71]	9.791	7.693	13.291	1.829	1.328
	DepthAnything [96]	4.105	3.680	6.130	0.262	0.613
	DepthAnything-ft	*2.584*	*2.401*	*4.167*	*0.125*	*0.417*
	ZoeDepthN-ft	**2.429**	**2.343**	**4.002**	**0.121**	**0.400**
Basketball	PatchFusion [49]	3.586	4.223	4.430	0.196	0.443
	Marigold [43]	3.276	4.188	4.396	0.195	0.440
	ZoeDepth [5]	2.898	3.556	3.732	0.140	0.373
	MiDaS [71]	1.715	2.519	2.637	0.0745	0.264
	DepthAnything [96]	0.725	1.582	1.653	0.029	0.165
	DepthAnything-ft	**0.691**	**1.341**	**1.401**	**0.020**	**0.140**
	ZoeDepthN-ft	*0.741*	*1.399*	*1.463*	*0.023*	*0.146*

criterion procedure introduced by Ranftl *et al.* [71]. However, we do not include any threshold metric such as δ_1, δ_2, and δ_3, since they measure the proportion of pixels where the ratio between the ground-truth value and the estimate falls below a specified threshold. Therefore, the threshold value is directly influenced by the depth distribution, which is subjective for relative depth. Hence, the metric may not hold meaningful significance in our case. Finally, since SoccerNet-Depth contains two sports with distinct depth distributions, we evaluate each sport separately.

Baselines. We evaluate five state-of-the-art monocular depth estimation methods. Two are fine-tuned, as training codes for the others were not available. **(i) MiDaS** [71] computes relative inverse depth, also called disparity. The innovation came through the introduction of a new training loss called *scale-and-shift invariant*, allowing to mix distinct datasets for training. Ever since the work was published, multiple derived models were proposed using different encoder backbones [6], transitioning from convolutional methods to vision transformers such as ViT [18]. **(ii) ZoeDepth** [5] reunites both relative and metric depth estimation to boost performances. Built on a more recent version of the DPT [70] encoder-decoder, the method starts by leveraging the MiDaS [71] relative depth estimation framework to obtain the relative depth map. After that, the result constitutes the input for an enhanced version of the LocalBins [4] module to obtain a final metric depth estimation. **(iii) Depth Anything** [96] is a foundation model that aims to predict depth accurately for any images in a broad range of scenarios. It uses the power and quantity of unlabeled data by automatically annotating them to provide both zero-shot relative and metric depth estimation. **(iv) Marigold** [43] is a latent diffusion model that has been fine-tuned on synthetic data and can also provide zero-shot generalization. Using Stable Diffusion [74] pre-trained and a fine-tuned U-Net, the method encodes the input into its latent code and concatenates it with the depth latent code learned during training. The result is passed at each denoising operations to the adapted version of the U-Net. **(v) PatchFusion** [49] enables accurate depth map prediction on high-resolution images. The framework consists in three distinct steps: the first one, the Coarse Network, loses details while gaining global awareness of the images, the second one, the Fine Network, splits the input into patches to understand all the fine details and, finally, a Guided Fusion Network with a Global-to-Local (G2L) module combines those results. Mechanisms are implemented during training and inference to maintain consistency among patches.

Implementation details. For all baseline methods, we first use the inference code to obtain predictions on our test set. For MiDaS [71], we use the best pre-trained model, called BEiT_512-L [6], that leverages a BEIT [3] backbone trained at a 512 × 512 resolution. To feed the data to the encoder, we resize the images while keeping the aspect ratio. For ZoeDepth [5], we investigated two pre-trained models, respectively pre-trained on NYUv2(N) and on a mix of KITTI [28] and NYUv2(NK). For both, we use the recommended MiDaS [71] backbone. Additionally, we fine-tune both pre-trained models on each sport of our dataset. As we have high-resolution data, we keep a small batch size (*i.e.*, 4), and fine-tune the models on a Tesla V100 GPU for 12 epochs. We denote the resulting models as *ZoeDepthN-*

ft and *ZoeDepthNK-ft*. For Depth Anything [96], we use the DINOv2 [67] encoder for feature extraction and the DPT [70] decoder. To obtain metric depth predictions, they follow the ZoeDepth [5] framework, replacing only the MiDaS [71] encoder by their ViT-L encoder. To perform inference using DepthAnything [96], we use their ViT-L encoder, as it performs best on almost all datasets. Additionally, we denote *DepthAnything-ft* the fine-tuned version of the method on our dataset. The training is performed with a batch size of 4 for 12 epochs on a modified version of ZoeDepth where the initial encoder has been substituted with the DepthAnything pre-trained *ViT-L* encoder. For Marigold [43], we keep the default inference settings and use the pre-trained weights on Hypersim and Virtual Kitti. Particularly, as the predictions are rescaled to the original resolution by the method, we maintain the 768 × 768 processing resolution. This resolution is optimal for Stable Diffusion [74], the model from which Marigold is derived. Finally, for PatchFusion [49], we use the weights pre-trained on the MVS-Synth [37] dataset. Our choice is motivated by the fact that this dataset contains images of outdoor scenes extracted from a video game at a resolution of 1920 × 1080. To enhance the predictions, we specify the input resolution, activate the reduction of patch artifacts, and keep the number of random added patches to 128.

Main Results. The performances of the five state-of-the-art models are presented in Table 2. First, in inference mode, it can be noted that similar rankings are observed across both sports. Without specific training tailored to our dataset, Depth Anything [96] displays the best performances across all metrics. This can be explained by its objective to provide state-of-the-art zero-shot relative depth estimation. Notably, it significantly outperforms MiDaS v3.1 [6], whose objective is similar. Conversely, the two diffusion-based models, Marigold and PatchFusion, are underperforming on unseen sports data. Finally, we show that fine-tuning ZoeDepth [5](N) and Depth Anything [96] on each sport leads to state-of-the-art performances. *Depth Anything-ft* is the best performer in basketball and *ZoeDepthN-ft* in football. Hence, fine-tuning from the NYUv2 pre-trained weights enables the most substantial improvements on football data. This can be explained by the fact that there are no existing sport datasets dedicated to monocular depth estimation. Thus, even methods like MiDaS or Depth Anything, which are trained on a greater quantity of data, are not used to the specific football domain. However, basketball resembles a more daily life scenario with a closer point of view and indoor scenes, leading to better zero-shot performance. Note that the metrics scales depend on the relative depth distribution and are thus not comparable between the sports.

Table 3. **Effect of pre-trained weights and generalization capabilities.** F- and B-ZoeDepth-ft stands for a model fine-tuned either on Football or on Basketball, with initial weights obtained from a pre-training on the NUYv2 (N) or KITTI (K) dataset.

		Football		**Basketball**	
		REL	RMSE	REL	RMSE
Algorithm	Pre-training	$\times 10^{-3}$	$\times 10^{-3}$	$\times 10^{-3}$	$\times 10^{-3}$
F-ZoeDepth-ft	No	2.705	3.078	3.244	4.034
F-ZoeDepth-ft	N	2.429	**2.343**	**1.876**	**2.850**
F-ZoeDepth-ft	N+K	**2.405**	2.488	2.131	2.943
B-ZoeDepth-ft	No	12.885	9.175	1.234	2.134
B-ZoeDepth-ft	N	**9.785**	**8.354**	**0.741**	1.399
B-ZoeDepth-ft	N+K	12.301	10.049	0.761	**1.396**

Ablation study. In the previous section, we fine-tuned ZoeDepth on each sport individually and evaluated its performance on the same sports. This section first proposes an analysis on the importance of the pre-trained weights when fine-tuning on sports-specific distributions for ZoeDepth and then provides an out-of-domain performance analysis. In Table 3, we show that, for both sports, the fine-tuned model without pre-training performs slightly worse than the pre-trained models. For basketball, fine-tuning pre-trained weights obtained by combining the KITTI and NYUv2 dataset improves the RMSE performance compared to NYUv2 alone, while for football, it is the opposite. Next, the out-of-domain performance analysis explores the inference performance on the sport that the model was not trained on. Table 3 shows that fine-tuned pre-trained models outperform models trained from scratch in predicting depth values from scenes of the other sport.

Qualitative Results. Figure 4 displays 4 frames extracted from distinct sequences for a qualitative analysis of the predictions of state-of-the-art models. The selected images illustrate various conditions as well as different viewpoints. It can be seen that, without fine-tuning, Depth Anything [96] performs the best across both sports, which is confirmed by Table 2. The model accurately reproduces the depth values of the field plane while maintaining precise border descriptions of the players. Let us note that it is critical to differentiate between the precision of depth estimations and the aesthetic quality of depth maps. As an example, PatchFusion can accurately represent the details and joints of the players but is struggling to predict the ground plan.

Moreover, we analyze the effect of fine-tuning ZoeDepth [5] on both sports in Figure 5. As can be seen, the predictions are significantly better when fine-tuning the model. Furthermore, we observe that the model has a deeper understanding of the camera angle, allowing it to better depict the notion of plane for the football field, as evidenced by the shades of color becoming darker with increasing ground truth depth values. For both sports, the players silhouette and objects are also more precise, showing that fine-tuning on our dataset leads to state-of-the-art performance for monocular depth estimation in sports.

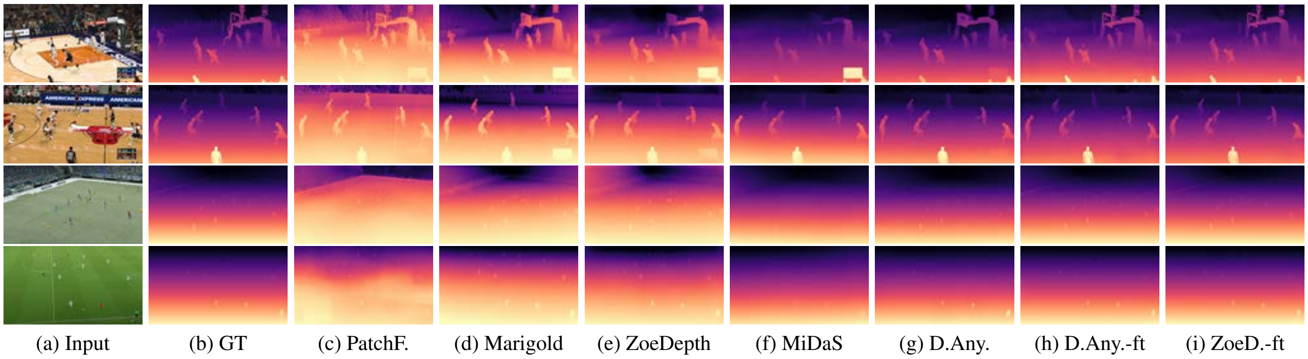

Figure 4. **Qualitative monocular depth estimation results.** Depth estimations predictions of state-of-the-art methods on our dataset. The first row shows a player holding onto the basketball hoop. The second row represents a close-up basketball image of a player initiating a play. The third row displays players relatively far from the camera and exhibits a shot at the goal in snowy conditions. The last row displays a football game with two players running for the ball near the camera. For each example, we display the (a) input RGB image, (b) ground-truth depth map, and (c-i) the predicted depth maps of each method.

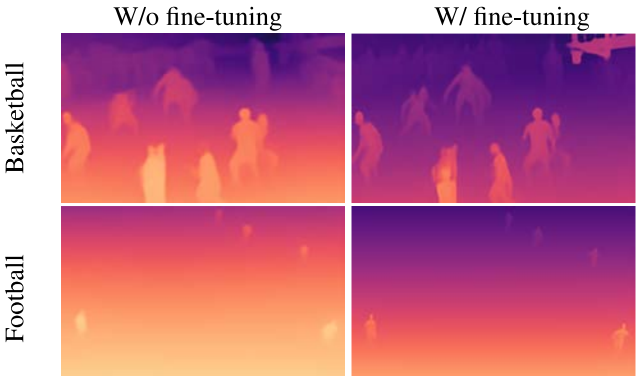

Figure 5. **ZoeDepthN-ft qualitative results.** Comparison of the predicted depth maps of ZoeDepth with and without fine-tuning on our dataset for each sport. Fine-tuning improves the objects and players borders, as well as the understanding of the field plane.

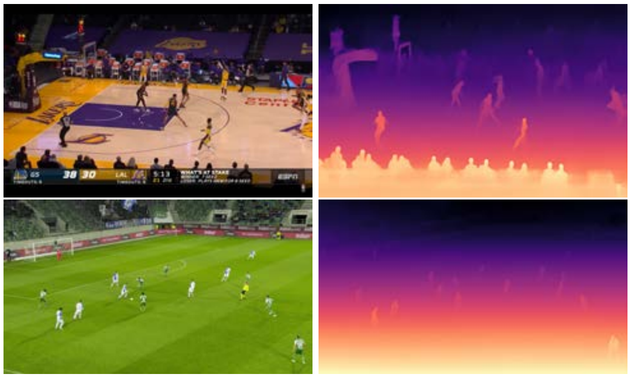

Figure 6. **Sim2real gap analysis.** Predicted depth maps of the ZoeDepthN-ft model on frames extracted from two real-world video sequences. The top row presents the estimation for an actual NBA match, while the bottom row features the prediction for a Swiss-League football match.

Finally, we provide a *Sim2Real* gap analysis in Figure 6. To do so, we apply our *ZoeDepthN-ft* model on real video sequences from an NBA match and a professional football game. It can be seen that the depth maps produced by our model show promising results as the model still understands the notion of field, demonstrated by the shades of color increasing with the further part of the field, while keeping the players silhouettes relatively well separated from the field.

5. Conclusion

We release the new SoccerNet-Depth dataset, which is the first scalable, synthetic dataset dedicated to monocular depth estimation in team sports videos. Our dataset is built by leveraging the high-quality sports simulations provided by two sport video games: NBA2K22 and Efootball. Particularly, we automatically extract video sequences and their associated depth maps, which can easily be adapted to other sport simulations. Furthermore, we benchmark five state-of-the-art methods on the monocular depth estimation task, highlighting their performance and fine-tune some of them on our dataset to establish a new state-of-the-art result on our dataset. We show that the models trained on synthetic data transfer well on real data, and provide an ablation study on pre-training and generalization. Finally, our work aims to encourage research in monocular depth estimation in sports through the release of future challenges.

Acknowledgment. A. Cioppa is funded by the F.R.S.-FNRS. This work was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 5850–5859, Waikoloa, HI, USA, Jan. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[2] Adrià Arbués Sangüesa, Adriàn Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3875–3884, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv*, abs/2106.08254, 2021. 6

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. LocalBins: Improving depth estimation by learning local distributions. *arXiv*, abs/2203.15132, 2022. 2, 6

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, abs/2302.12288, 2023. 2, 6, 7

[6] Reiner Birkl, Diana Wofk, and Matthias Müller. MiDaS v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv*, abs/2307.14460, 2023. 6, 7

[7] Matthias Boeker and Cise Midoglu. Soccer athlete data visualization and analysis with an interactive dashboard. In *Int. Conf. Multimedia Retr.*, volume 13833 of *Lect. Notes Comput. Sci.*, pages 565–576. Springer Int. Publ., 2023. 3

[8] Bruno Cabado, Anthony Cioppa, Silvio Giancola, Andrés Villa, Bertha Guijarro-Berdiñas, Emilio Padrón, Bernard Ghanem, and Marc Van Droogenbroeck. Beyond the Premier: Assessing action spotting transfer capability across diverse domains. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, Seattle, WA, USA, Jun. 2024. 3

[9] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Sci. Data*, 9(1):1–9, Jun. 2022. 3

[10] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13123–13133, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[11] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 4532–4541, Nashville, TN, USA, Jun. 2021. 3

[12] Anthony Cioppa, Adrien Deliege, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 2505–2514, Long Beach, CA, USA, Jun. 2019. Inst. Electr. Electron. Eng. (IEEE). 3

[13] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 1846–1855, Salt Lake City, UT, USA, Jun. 2018. 3

[14] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3490–3501, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 3

[15] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmy, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuha, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Hérault, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yanbiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huangi, and Ziyu Meng. SoccerNet 2023 challenges results. *arXiv*, abs/2309.06006, 2023. 3

[16] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 4508–4519, Nashville, TN, USA, Jun. 2021. 3

[17] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers. *arXiv*, abs/2309.01270, 2023. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, abs/2010.11929, 2020. 2, 6

[19] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 2366–2374, 2014. 2, 5

[20] Baba Fakhar, Hamidreza Rashidy Kanan, and Alireza Behrad. Event detection in soccer videos using unsupervised learning of spatio-temporal features based on pooled spatial pyramid model. *Multimedia Tools Appl.*, 78(12):16995–17025, Jun. 2019. 3

[21] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4008–4017, Nashville, TN, USA, Jun. 2021. Inst. Electr. Electron. Eng. (IEEE). 2

[22] Michaël Fonder. *Reliable Monocular Depth Estimation for Unmanned Aerial Vehicles*. PhD thesis, University of Liège, Belgium, Jul. 2023. 3

[23] Michaël Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), UAVision*, pages 553–562, Long Beach, CA, USA, Jun. 2019. Inst. Electr. Electron. Eng. (IEEE). 3

[24] Xubo Fu, Kun Zhang, Changgang Wang, and Chao Fan. Multiple player tracking in basketball court videos. *J. Real-Time Image Process.*, 17(6):1811–1828, Apr. 2020. 3

[25] Xin Gao, Xusheng Liu, Taotao Yang, Guilin Deng, Hao Peng, Qiaosong Zhang, Hai Li, and Junhui Liu. Automatic key moment extraction and highlights generation based on comprehensive soccer video understanding. In *IEEE Int. Conf. Multimedia Expo Work. (ICMEW)*, pages 1–6, London, Engl., Jul. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[26] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 9912 of *Lect. Notes Comput. Sci.*, pages 740–756. Springer Int. Publ., 2016. 2

[27] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri Kshatri, and Paal Halvorsen. Assisting soccer game summarization via audio intensity analysis of game highlights. In *IOE Graduate Conference*, volume 12, pages 25–32, Oct. 2022. 3

[28] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.*, 32(11):1231–1237, Aug. 2013. 3, 6

[29] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 1792–179210, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 3

[30] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 challenges results. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 75–86, Lisbon, Port., Oct. 2022. ACM. 3

[31] Silvio Giancola, Matteo Valenti, and Remo Sala. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springerbriefs Comput. Sci. Springer Int. Publ., 2018. 1

[32] GitHub. Python-ImageSearch. `https://github.com/drov0/python-imagesearch`. 4

[33] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, second edition, 2004. 1

[34] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. VARS: Video assistant referee system for automated soccer decision making from multiple views. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5086–5097, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 3

[35] Jan Held, Hani Itani, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. X-vars: Introducing explainability in football refereeing with multimodal large language models. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, Seattle, WA, USA, Jun. 2024. 3

[36] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. *arXiv*, abs/2207.10213, 2022. 3

[37] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *IEEE/CVF Conf. Comput. Vis. Pattern

Recognit. (CVPR), pages 2821–2830, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2, 3, 7

[38] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 9–18, Seattle, WA, USA, Oct. 2020. 3

[39] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. SportsPose - a dynamic 3D sports pose dataset. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5219–5228, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 3

[40] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 23–29, Ottawa, Ontario, Can., Oct. 2023. ACM. 3

[41] Ali Karimi, Ramin Toosi, and Mohammad Ali Akhaee. Soccer event detection using deep learning. *arXiv*, abs/2102.04331, 2021. 3

[42] Baldur Karlsson. RenderDoc. https://renderdoc.org, 2018. 3, 4

[43] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv*, abs/2312.02145, 2023. 2, 6, 7

[44] Muhammad Zeeshan Khan, Summra Saleem, Muhammad A. Hassan, and Muhammad Usman Ghanni Khan. Learning deep C3D features for soccer video event detection. In *Int. Conf. Emerg. Technol. (ICET)*, pages 1–6, Islamabad, Pakistan, Nov. 2018. 3

[45] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Trans. Image Process.*, 27(8):4131–4144, Aug. 2018. 2, 3

[46] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimarães, and Pietro Perona. Text-image alignment for diffusion-based perception. *arXiv*, abs/2310.00031, 2023. 2

[47] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv*, abs/1907.10326, 2019. 2

[48] Chuxuan Li, Ran Yi, Saba Ghazanfar Ali, Lizhuang Ma, Enhua Wu, Jihong Wang, Lijuan Mao, and Bin Sheng. RADepthNet: Reflectance-aware monocular depth estimation. *Virtual Reality & Intelligent Hardware*, 4(5):418–431, Oct. 2022. 3, 4

[49] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. PatchFusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. *arXiv*, abs/2312.02284, 2023. 2, 6, 7

[50] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. DepthFormer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, Sept. 2023. 2

[51] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2041–2050, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2, 3

[52] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. BinsFormer: Revisiting adaptive bins for monocular depth estimation. *arXiv*, abs/2204.00987, 2022. 2

[53] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2024–2039, Oct. 2016. 2

[54] Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. Learning selective mutual attention and contrast for RGB-d saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):9026–9042, Dec. 2022. 2

[55] Katja Ludwig, Julian Lorenz, Robin Schön, and Rainer Lienhart. All keypoints you need: Detecting arbitrary keypoints on the body of triple, high, and long jump athletes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5179–5187, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 3

[56] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *arXiv*, abs/2004.15021, 2020. 2

[57] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. A universal protocol to benchmark camera calibration for sports. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, Seattle, WA, USA, Jun. 2024. 3

[58] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3460–3470, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 3

[59] Amir M. Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 103–112, Ottawa, Ontario, Can., Oct. 2023. ACM. 3

[60] Cise Midoglu, Steven Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen. MMSys'22 grand challenge on AI-based video production for soccer. In *ACM Multimedia Systems Conference (MMSys)*, pages 1–6, Athlone, Ireland, Jun. 2022. 3

[61] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5074–5085, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 3

[62] Thomas B. Moeslund, Graham Thomas, and Adrian Hilton. *Computer vision in sports*. Springer, 2014. 3

[63] Matthias Müller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. Driving policy transfer via modularity

and abstraction. In *Conference on Robot Learning (CoRL)*, pages 1–15, Zürich, Switzerland, Oct. 2018. 3

[64] Banoth Thulasya Naik, Mohammad Farukh Hashmi, Neeraj Dhanraj Bokde, and Zaher Mundher Yaseen. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Appl. Sci.*, 12(9):1–49, Apr. 2022. 3

[65] Danish Nazir, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. SemAttNet: Towards attention-based semantic aware guided depth completion. *arXiv*, abs/2204.13635, 2022. 2

[66] NVidia. NVIDIA Nsight Graphics. https://developer.nvidia.com/nsight-graphics. 4

[67] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and et al. DINOv2: Learning robust visual features without supervision. *arXiv*, abs/2304.07193, 2023. 7

[68] Nithin Raghavan, Punarjay Chakravarty, and Shubham Shrivastava. Sim2Real for self-supervised monocular depth and segmentation. *arXiv*, abs/2012.00238, 2020. 3

[69] Aakash Rajpal, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, and Sunil Jaiswal. High-resolution synthetic RGB-D datasets for monocular depth estimation. *arXiv*, abs/2305.01732, 2023. 3

[70] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 12159–12168, Montréal, Can., Oct. 2021. Inst. Electr. Electron. Eng. (IEEE). 2, 6, 7

[71] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637, Mar. 2022. 2, 6, 7

[72] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4738–4747, Salt Lake City, UT, USA, Jun. 2018. 3, 4, 5

[73] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 9906 of *Lect. Notes Comput. Sci.*, pages 102–118. Springer Int. Publ., 2016. 3, 4

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10684–1695, New Orleans, LA, USA, Jun. 2022. 6, 7

[75] Himangi Saraogi, Rahul Anand Sharma, and Vijay Kumar. Event recognition in broadcast soccer videos. In *Indian Conf. Comput. Vis. Graph. Image Process.*, page 1–7. ACM, Dec. 2016. 3

[76] Mehdi Houshmand Sarkhoosh, Sayed Mohammad Majidi Dorcheh, Cise Midoglu, Saeed Shafiee Sabet, Tomas Kupka, Dag Johansen, Michael A. Riegler, and Pål Halvorsen. AI-based cropping of soccer videos for different social media representations. In *Int. Conf. Multimedia Retr.*, volume 14557 of *Lect. Notes Comput. Sci.*, pages 279–287. Springer Nat. Switz., 2024. 3

[77] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 18, pages 1–8, Vancouver, British Columbia, Canada, Dec. 2005. Curran Assoc. Inc. 2

[78] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, volume 8753 of *Lect. Notes Comput. Sci.*, pages 31–42. Springer, 2014. 3

[79] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3568–3578, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 3

[80] Karolina Seweryn, Gabriel Cheć, Szymon Łukasik, and Anna Wróblewska. Improving object detection quality in football through super-resolution techniques. *arXiv*, abs/2402.00163, 2024. 3

[81] Karolina Seweryn, Anna Wróblewska, and Szymon Łukasik. Survey of action recognition, spotting and spatio-temporal localization in soccer – current trends and research perspectives. *arXiv*, abs/2309.12067, 2023. 3

[82] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 7576 of *Lect. Notes Comput. Sci.*, pages 746–760, Firenze, Italy, Oct. 2012. 3

[83] João V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2796–2800, Bordeaux, France, Oct. 2022. Inst. Electr. Electron. Eng. (IEEE). 3

[84] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, Seattle, WA, USA, Jun. 2024. 3

[85] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 567–576, Boston, MA, USA, Jun. 2015. Inst. Electr. Electron. Eng. (IEEE). 3

[86] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: current applications and research topics. *Comput. Vis. Image Underst.*, 159:3–18, Jun. 2017. 3

[87] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-nets: Stereo mixture density networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8938–8948, Nashville, TN, USA, Jun. 2021. Inst. Electr. Electron. Eng. (IEEE). 3

[88] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 1–8, Lisbon, Port., Oct. 2022. ACM. 3

[89] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3480–3489, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 3

[90] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and et al. DIODE: A Dense Indoor and Outdoor DEpth dataset. *arXiv*, abs/1908.00463, 2019. 3

[91] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, abs/1706.03762, 2017. 2

[92] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. TTNet: Real-time temporal and spatial video analysis of table tennis. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3866–3874, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[93] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. *arXiv*, abs/1904.11112, 2019. 2

[94] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. *arXiv*, abs/2112.02306, 2021. 3

[95] Lifang Wu, Zhou Yang, Qi Wang, Meng Jian, Boxuan Zhao, Junchi Yan, and Chang Wen Chen. Fusing motion patterns and key visual information for semantic event recognition in basketball videos. *arXiv*, abs/2007.06288, 2020. 3

[96] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024. 2, 6, 7

[97] Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. MAMo: Leveraging memory and attention for monocular video depth estimation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 8720–8730, Paris, Fr., Oct. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[98] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3D scene shape from a single monocular image. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 6480–6494, 2023. 2

[99] Young Yoon, Heesu Hwang, Yongjun Choi, Minbeom Joo, Hyeyoon Oh, Insun Park, Keon-Hee Lee, and Jin-Ha Hwang. Analyzing basketball movements and pass relationships using realtime object tracking techniques based on deep learning. *IEEE Access*, 7:56564–56576, 2019. 3

[100] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected CRFs for monocular depth estimation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3906–3915, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[101] Weichen Zhang, Zhiguang Liu, Liuyang Zhou, Howard Leung, and Antoni B. Chan. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. *Image Vis. Comput.*, 61:22–39, May 2017. 3, 5

[102] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single RGB-D image. *arXiv*, abs/1803.09326, 2018. 2

[103] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 5706–5716, Paris, Fr., Oct. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[104] Luyang Zhu, Konstantinos Rematas, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Reconstructing NBA players. *arXiv*, abs/2007.13303, 2020. 3, 4