A Novel Synthetic Dataset for Broadcast Motorsports Scene Understanding

A Novel Synthetic Dataset for Broadcast Motorsports Scene Understanding

Luca Francesco Rossi [1,2], Andrea Sanna [1], Federico Manuri [1]
[1]Department of Control and Computer Engineering
Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129, Torino, TO, Italy
e-mail: {lucafrancesco.rossi | andrea.sanna | federico.manuri}@polito.it

Mattia Donna Bianco[2]
[2]netventure R&D S.r.l.
Software Engineer
Via della Consolata, 1\bis, 10122, Torino, TO, Italy
e-mail: m.donnabianco@netventure.tv

***Abstract*—The paper introduces a foundational approach to motorsports scene understanding by investigating the role of synthetic data generation in advancing scene understanding for high-speed broadcast scenarios. Utilizing the CARLA (Car Learning to Act) simulation environment, the study constructs a high-fidelity dataset incorporating diverse lighting conditions, occlusions, and dynamic camera perspectives to enhance model generalization. A multi-stage data refinement pipeline is introduced to mitigate the impact of extreme occlusions and irrelevant samples while preserving the complexity of real-world challenges. Possible applications include 3D real-world understanding from a single monocular 2D image, which could open up interesting possibilities for augmented reality in broadcast media by allowing seamless integration of virtual elements, interactive graphics and dynamic visual effects, enhancing storytelling, audience engagement, and production flexibility. The efficacy of the dataset is further evaluated via transfer learning to the real-world domain, with the model pretrained on synthetic data demonstrating a significantly superior performance compared to its counterpart.**

Keywords-computer vision; augmented reality; synthetic data generation; transfer learning.

I. INTRODUCTION

Computer Vision (CV) algorithms based on Artificial Intelligence (AI) are revolutionizing the sports industry, offering advanced analytical capabilities that enhance performance evaluation, officiating accuracy, and fan engagement. By leveraging AI-driven techniques, such as player tracking, ball trajectory estimation and action recognition, these algorithms provide real-time insights that were previously considered unattainable [1]. Coaches and analysts can use this technology to refine strategies, optimize training regimens and prevent injuries by closely monitoring player movements and biomechanics. Referees benefit from automated decision-making tools that minimize human error and ensure fair play, while broadcasters utilize computer vision to generate augmented replays, statistical overlays and personalized viewing experiences. Yet, while a noticeable surge of interest towards these techniques has been observed in a multitude of sports [2]–[4], the specific field of motorsports has traditionally been regarded as exclusively linked to industrial applications, with minimal to no scholarly research available in the literature addressing the analysis and comprehension of event dynamics and racing scenarios.

Therefore, this study was undertaken to address this gap by introducing a synthetic dataset that includes 3D information on ABB FIA Formula E Gen3 racing car models in urban environments, considering as underlying objective to establish a foundation for advancing research in motorsports scene understanding.

The structure of the paper is as follows: Section II reviews relevant literature pertinent to the present study; Section III outlines the methodology and technical details employed in constructing the synthetic dataset, discussing both its advantages and limitations; Section IV assesses the reliability of the dataset by examining synthetic-to-synthetic and synthetic-to-real performance, exploring potential applications of the work; and Section V concludes the paper, summarizing the findings and suggesting avenues for future research.

II. RELATED WORK

The increasing popularity of AI, particularly in subfields like Machine Learning (ML) and Deep Learning (DL), has led to a significant challenge in the limited size (or lack) of training datasets. This limitation is primarily due to high workloads and privacy concerns, which hinder the model's ability to generalize effectively [5]. Synthetic Data Generation (SDG) arose as a viable solution to address such an issue: by generating artificial data and labels that closely emulate authentic samples, it alleviates constraints imposed by traditional datasets. This approach proves highly valuable when real data is insufficient, costly to label or exhibits biased distributions, and its advantages go beyond cost reduction, contributing to reduced computational time and addressing bias in data distribution. Eventually, synthetic data can also be generated on the fly during training, eliminating the need for storage, and can be made to be as photorealistic as possible, allowing models to transfer from synthetic training sets to real test sets.

A. Synthetic data generation for sports

Cerqueira and Kenwright [6] introduced a novel approach to CV-based feature extraction in football by leveraging entirely synthetic training data. Differently from conventional machine learning models in sports analytics that typically depend on real-world images, such a study investigates instead the feasibility of training machine learning models exclusively on synthetic datasets generated through computer graphics, with the objective of minimizing the domain gap between synthetic and real-world data [7]. By generating high-fidelity, labeled synthetic images of football matches and by incorporating a diverse range of viewpoints, lighting conditions, occlusions, and visual artifacts, the authors demonstrate that models trained exclusively on synthetic data can generalize effectively to real-world football imagery, accurately identifying pitch markers and player positions. The study validates the potential of synthetic data to address key limitations of real-world datasets, demonstrating its efficacy in the application of synthetic data for sports analytics.

Bhargavi et al. [8] demonstrated that the integration of synthetic data with lightweight deep learning models can achieve state-of-the-art results in jersey number identification while minimizing the need for extensive manual annotations or large-scale datasets. The proposed method involves an initial step of detecting and segmenting players from video frames using a pretrained person detection model [9]. Subsequently, a human pose estimation model [10] is employed to localize jersey numbers by identifying torso key points, thereby obviating the need for manual annotation of bounding boxes. Given the constraints of real-world datasets in terms of sample size and class imbalance, the study introduces two synthetic datasets – Simple2D and Complex2D.

Qin et al. [11] presented SoccerSynth-Detection, a novel synthetic dataset specifically designed for soccer player detection, addressing the limitations of existing real-world datasets, such as SoccerNet-Tracking [12] and SportsMoT [13]. To construct the dataset, the authors augmented a previously developed soccer stadium simulator by integrating a central camera with configurations derived from real-world match footage. They employed assets from the Unreal Engine Marketplace to model player appearances and animations, while movement logic was implemented through AI-controlled Behavior Trees. The simulation environment was further enhanced by incorporating dynamic lighting, randomized textures and motion blur, thereby mitigating the domain gap between synthetic and real-world data to improve model generalization. In the transfer learning experiment, a model trained on SoccerSynth-Detection was evaluated against real-world datasets. While it exhibited a slight reduction in AP50 performance [14] compared to real datasets, it demonstrated superior results in more stringent detection settings (mAP50-95), particularly in handling motion blur, suggesting that the synthetic dataset can either match or surpass real datasets under specific conditions.

B. Scene understanding in motorsports

Boiarov et al. [15] presented RaceLens, a novel application that utilizes deep learning and computer vision models to automatically analyze racing photos. It is designed to maximize the potential of racing photographs by identifying and interpreting crucial elements in the images, such as detecting racing cars, recognizing car numbers, and detecting and quantifying car details. The proposed method employs a Metric Learning [16] approach to tackle the task, where the main encoder model takes a 3-channel image as input and outputs a 1-D vector representing the color scheme of the car in the image. The embeddings are trained to be closer to each other for images of the same class and farther apart for different classes, using a triplet loss and a fully connected layer with cross-entropy loss. During the inference phase, clusters can be created using the embeddings, and the so-called Car Number Recognition Model [17] is utilized to assign the corresponding team names to the clusters. The method allows for clustering of images based on color scheme and uses the Car Number Recognition Model to assign team names to the clusters, enabling the affiliation of cars with their respective teams. It has been deployed for NASCAR teams and has processed over 200 race events, with an average of 7000 photos per event, and has achieved high accuracy in its analysis, with an average percent of photos without cars being less than 1%. The framework uses a combination of models, including Keypoint R-CNN with ResNet-50 backbone [18], and has been evaluated using COCO metrics, achieving high average precision and recall.

Tyo et al. [19] presented the Racer Number Dataset (RnD), a novel and challenging dataset aimed at advancing research in Optical Character Recognition (OCR) within the domain of off-road motorsports. The dataset comprises 2,411 images collected from professional motorsports photographers across 50 distinct off-road competitions, encompassing a total of 5,578 manually annotated bounding boxes that delineate visible motorcycle racer numbers. These images present a range of conditions that pose significant challenges to OCR systems, including occlusions caused by mud, motion blur, glare, complex backgrounds, and non-standardized fonts. To assess the efficacy of contemporary OCR techniques in this domain, the authors conducted a benchmarking study using two state-of-the-art OCR models [20][21] – both in their pre-trained configurations and after fine-tuning on the RnD dataset – underscoring the necessity for domain-specific OCR techniques that are robust to extreme visual conditions, particularly in the context of motorsports.

Tyo et al. [22] further extended their work by presenting the Muddy Racer re-iDentification Dataset (MUDD), similarly designed to advance research in computer vision applications for off-road motorsports. The MUDD dataset consists of 3,906 images depicting 150 distinct riders across ten competitions, specifically curated for the task of rider re-identification (ReID). In line with their previous findings, empirical results underscore the necessity for domain-specific adaptations to enhance OCR and ReID performance in real-world motorsports

applications, with benchmark evaluations conducted using state-of-the-art OCR and ReID models [23] revealing that existing pre-trained models perform inadequately in such a domain. Promising results are nonetheless retrieved via a Contrastive Multiple Instance Learning (CMIL) framework [24] which introduces a new formulation that enables contrastive learning at the bag level: instead of focusing on individual image representations, CMIL optimizes entire bag representations, encouraging similar bags to have closer representations while pushing apart dissimilar ones.

III. METHOD

CARLA (Car Learning to Act) is an open-source urban driving simulator specifically designed to facilitate research in autonomous driving [25]. Developed through a collaboration between Intel Labs, the Toyota Research Institute, and the Computer Vision Center in Barcelona, it provides a sophisticated simulation environment for the development, testing, and validation of autonomous driving systems. A distinguishing characteristic of CARLA is its fully open-source nature, which includes an extensive collection of freely available digital assets, encompassing urban layouts, vehicles, pedestrians, and environmental elements. The platform enables the customization of sensor configurations, incorporating RGB cameras, depth sensors, and semantic segmentation, thereby allowing for comprehensive experimentation with perception systems. Furthermore, CARLA offers a dynamic simulation environment, supporting variable weather conditions, lighting scenarios, and traffic situations involving both autonomous and non-player vehicles as well as pedestrians, thus ensuring a high degree of realism and adaptability. The interested reader is recommended to discover more about CARLA in [26].

The rationale behind this work is that pose estimation for rigid bodies provides a fundamental approach to inferring three-dimensional (3D) spatial relationships from two-dimensional (2D) image data, enabling a deeper understanding of object orientation and motion within a scene. By leveraging keypoint detection and geometric transformations, pose estimation algorithms recover essential structural information, with consequent mapping of 2D projections to 3D coordinates through Perspective-n-Point (PnP) methods [27] leading to spatial understanding. Considering broadcasting applications, a 6-keypoints pose representation for race cars is proposed, selecting those keypoints that remain predominantly visible under typical viewing conditions. These keypoints include the four wheels, the top of the front wing, and the camera mount, ensuring robust and consistent pose estimation in dynamic racing environments.

A. Dataset preparation

Since originally developed for autonomous driving scenarios, the first extension required for modeling realistic motorsports images in CARLA consists in decoupling recording sensors and cameras from the ego vehicle to simulate possible broadcasting-level panoramic views. This is obtained by synchronously moving all cameras and sensors from one vehicle to the other at each world tick by applying a geometric transformation $\mathbb{T}_\tau$ to the camera position c_τ and orientation ρ_τ at tick time τ, i.e.,

$$\begin{bmatrix} c \\ \rho \end{bmatrix}_{\tau+1} = \mathbb{T}_\tau \left(\begin{bmatrix} c \\ \rho \end{bmatrix}_{\tau} \right) \tag{1}$$

where the geometric transformation is computed in such a way that

$$\mathbb{T}_\tau = \bar{\mathbb{T}}_\tau + \mathbb{X}_\tau \tag{2}$$

with $\bar{\mathbb{T}}_\tau$ being the transformation that would precisely bring the camera to point towards the chosen vehicle, and $\mathbb{X}_\tau$ is the instantiation at tick time τ of possible noisy operating camera movements observable in the real setting, such as zoom in, zoom out or random rotations that force the camera to drift away from always having the target exactly at the center of the image. Qualitatively, the optimal results were achieved by configuring the camera's field of view to 90° and letting $\bar{\mathbb{T}}_\tau$ positioning it along a circular trajectory with a radius of 7.5 meters, centered on the target vehicle's position. The camera was placed at a height of 3.5 meters above ground level and oriented directly toward the vehicle, irrespective of the specific point along the circumference. The transformation noise $\mathbb{X}_\tau$ is introduced to simulate zooming effects by applying a random shift within the interval $[-2, 2]$ meters to both the radius and height. Additionally, imprecision in camera orientation is incorporated by applying random variations in the yaw and pitch angles within the interval $[-15°, 15°]$ and in the roll angle within the interval $[-5°, 5°]$.

Three distinct lighting conditions – noon, sunset, and evening – were considered, with $1,500$ images generated for each setting, resulting in a total of $4,500$ frames at a resolution of 1920×1080 pixels.

B. Data refinement

One of the primary advantages of utilizing synthetic data is the availability of precise ground-truth information regarding the 3D spatial distribution at every stage. However, data processed by neural networks typically resides within the camera coordinate system, necessitating a transformation pipeline between the real 3D world and its abstracted 2D sensor representation. Figure 1 provides a step-by-step visual representation of the proposed approach discussed here:

A) the RGB frame is captured at world tick time τ;
B) real-world coordinates are employed to project the 3D bounding box coordinates onto the image plane;
C) an initial estimation of the 2D bounding boxes is obtained by identifying the extreme coordinates of the original 3D bounding boxes;
D) these preliminary 2D bounding boxes are further refined by computing the minimal enclosing rectangle of the vehicle's mask convex hull, extracted through semantic segmentation;
E) pose and keypoint visibility are filtered using a point cloud generated by a LiDAR sensor, distinguishing visible points from occluded ones relative to the camera perspective;

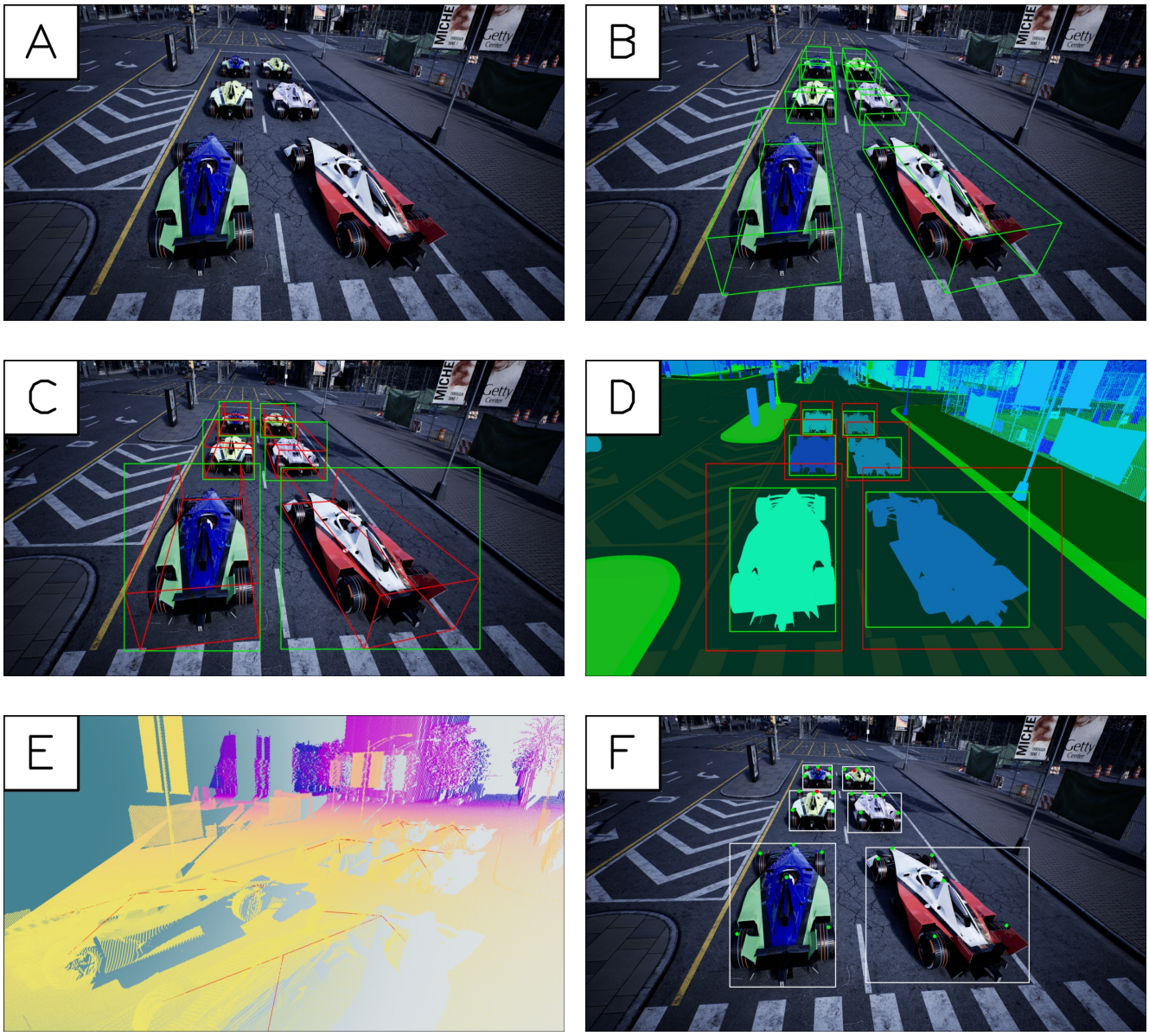

Figure 1. Qualitative visualization of the iterative SDG pipeline in CARLA.

F) the final dataset consists of the refined 2D bounding boxes, along with the corresponding poses and keypoints' visibility.

To minimize the presence of pure background images in the dataset, frames in which the relevant pixel area—defined as the number of pixels labeled by semantic segmentation as belonging to the actor of interest—was less than 1% of the total image size were automatically discarded during the data generation pipeline. Likewise, a maximum distance of 150 meters between an actor and the camera was established as a threshold for determining its relevance. A final criterion for actor inclusion was based on the ratio between its pixel area and the size of its refined 2D bounding box: if this ratio fell below 10%, the actor was automatically excluded by the SDG pipeline. Despite being qualitatively determined, these thresholds were kept very loose in order to just remove noisy information, such as complete occlusions or extreme aspect ratios and bounding box sizes. With respect to keypoint visibility, a threshold of 0.3 meters was established. More in detail, a spherical region with this radius, centered at the actual 3D location of the keypoint, is defined in the point cloud: if no other point is detected within such neighborhood, the keypoint is considered as not visible.

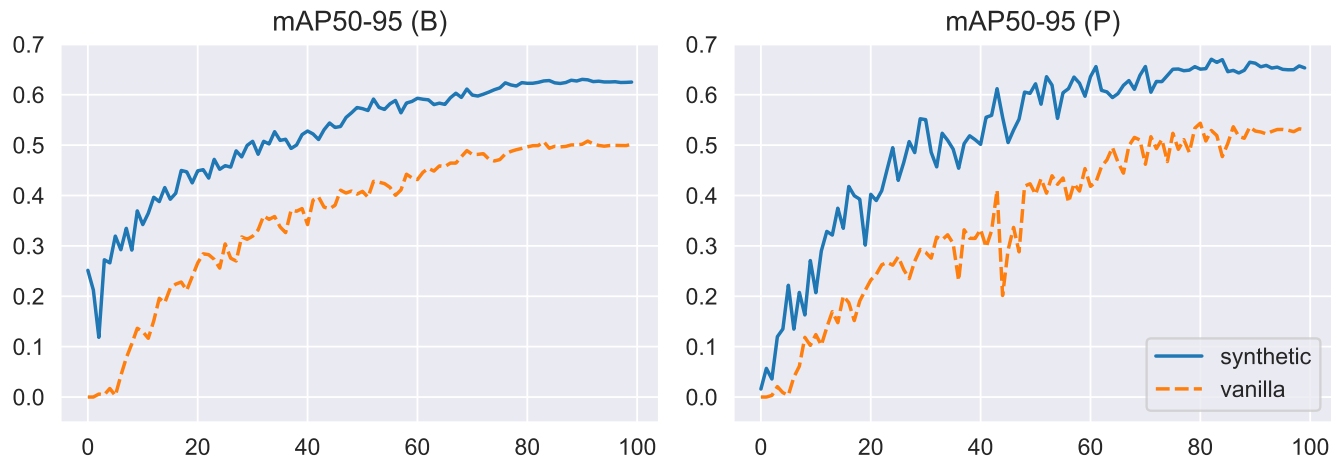

Figure 2. Validation mAP50-95 box (B) and pose (P) metrics gap on real data with and without transfer learning from the proposed synthetic dataset.

C. Limitations

Given the significant variability in camera orientations and the impact of partial environmental occlusions, certain deficiencies of such a fully automated approach must be acknowledged. First, when refining the 2D bounding box via semantic segmentation, it is assumed that the convex hull of the most frequently labeled vehicle pixels corresponds to the actor of interest. This assumption is generally robust, but some care has to be taken in those rare cases of strong occlusions among vehicles, where mis-classifications might occur.

Second, in order to build the point cloud for visibility computation, a static ray-casting should ideally be set from the camera perspective. This is not exactly what is done by CARLA LiDAR sensor [26], which behaves as a solid approximation but few (ideally not visible) points might still be present in the cloud. For such a reason, a strategy of hidden point removal [28] is implemented: by identifying the points located on the convex hull of a transformed cloud, visibility can be determined without neither the need for surface reconstruction nor normal estimation [29].

IV. Evaluation and Discussion

This section presents empirical results obtained using the dataset introduced thus far. Given the inherently fast-moving dynamics of motorsports scenarios, the You Only Look Once (YOLO) framework [30][31] has been chosen to simulate inference under real-time constraints. From the entire dataset, 3,150 images (70%) were allocated for training, 900 images (20%) for validation, and 450 images (10%) for testing.

More in detail, the whole training process was executed on a single NVIDIA Tesla V100 SXM2 GPU (32 GB, 5120 CUDA cores), inside a Python 3.7.7 environment with PyTorch 1.31.1 for CUDA 11.6. A YOLOv8x model was trained for 100 epochs with mixed precision on batches of eight 1280×1280 resized images. A cosine scheduler was set, progressively reducing the learning rate from its initial value of 1e-4 to a hundredth of it.

Table I and Table II highlight the best COCO metric values – on both validation and test splits – for vehicle bounding box detection and keypoints pose estimation, respectively. Concerning such results, a peculiar disparity between precision and recall is evident, likely attributable to a non-negligible presence of false negatives, as inferred from the high precision value. The suboptimal performance observed on the dataset may partially stem from the loosely-defined exclusion criteria applied during its construction. If the exclusion process fails to properly eliminate all borderline cases, i.e., ground truth vehicles under challenging occlusions or strong out-of-frame – the overall recall metric may as results be negatively impacted by the dataset's compromised quality rather than the inherent limitations of the model itself. Yet, while such exclusion criteria may introduce very challenging scenarios for the model, this characteristic can be seen as an advantage rather than a deficiency. By retaining most borderline cases, the dataset better reflects the complexities of real-world scenarios, where perfect visibility and ideal conditions are rarely guaranteed. Instead of filtering out these challenging instances, their inclusion provides a more comprehensive evaluation of the model's robustness and generalization ability. This approach ensures that the model is trained and tested on a diverse range of conditions, ultimately possibly leading to improved performance in practical broadcasting applications where imperfect data is the norm.

TABLE I. SYNTHETIC DATASET METRICS FOR BBOX DETECTION.

Split	P_B	R_B	$mAP50_B$	$mAP50\text{-}95_B$
Validation	0.952	0.793	0.897	0.793
Test	0.962	0.773	0.884	0.784

TABLE II. SYNTHETIC DATASET METRICS FOR POSE ESTIMATION.

Split	P_P	R_P	$mAP50_P$	$mAP50\text{-}95_P$
Validation	0.924	0.728	0.842	0.816
Test	0.918	0.716	0.827	0.795

Figure 3. Qualitative illustration of dataset samples with 3D space reconstruction after PnP computation.

A. *Synthetic to real adaptation*

The disparity in performance between synthetic and real-domain data remains a subject of ongoing discussion within the research community [32]. Reducing this gap is crucial to enhance both the reliance on and the applicability of synthetic data [33]. Given that models trained exclusively on synthetic data continue to exhibit suboptimal performance when applied to real-world scenarios [34], this study conducts a qualitative evaluation of the proposed dataset via transfer learning from the synthetic domain to the real one, with the objective of determining whether this approach leads to any improvement in model performance.

For this purpose, a proprietary dataset consisting of broadcast images from the 2023-2024 ABB FIA Formula E World Championship has been assembled from official broadcast racing highlights: 293 training frames from the Mexico City ePrix and 42 validation frames from the Portland ePrix were provided to five independent annotators to generate manually-labeled ground-truth annotations for bounding boxes and pose keypoints, including visibility information.

In accordance with the previously described experimental setup, two distinct YOLOv8x models – one pretrained on the synthetic dataset and the other initialized from scratch – were trained on the real data. Figure 2 presents the evolution of

the mAP50-95 metric over 100 epochs for both bounding box detection (on the left) and keypoints pose estimation (on the right) for the two models, emphasizing the disparity between the two performance trends. The results clearly illustrate the impact of synthetic pretraining, with the pretrained model demonstrating an improvement of 24.56% in mAP50-95 for bounding box detection and 23.08% for keypoints pose estimation, compared to its "vanilla" counterpart.

Table III offers a detailed summary of key statistics, providing an overview of the performance improvements observed across all training epochs.

TABLE III. RELEVANT STATISTICS CONCERNING mAP50-95 GAP.

Task	*AVG*	*STD*	*MIN*	*25%*	*50%*	*75%*	*MAX*
Box	0.161	0.042	0.113	0.127	0.149	0.183	0.317
Pose	0.164	0.056	0.016	0.128	0.152	0.205	0.354

B. PnP computation

The PnP problem consists in solving for the rotation and translation that minimizes the reprojection error from 3D-2D point correspondences. By reverse engineering the well-known problem [35]

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3}$$

one is therefore able to abstract the real-world 3D representation from the given 2D frame. Figure 3 qualitatively illustrates some dataset samples with corresponding 3D understanding reconstruction.

When transposed to the real setting, this approach could open up a whole set of opportunities: for example, understanding the 3D world from a single 2D image unlocks transformative possibilities for Augmented Reality (AR) in broadcast TV, enhancing storytelling, audience engagement, and real-time visual effects. By reconstructing 3D scenes from standard camera feeds, broadcasters can seamlessly integrate virtual objects, dynamic graphics, and interactive overlays into live or pre-recorded footage without requiring complex depth-sensing equipment. Real-time depth estimation also facilitates more natural occlusion handling, ensuring AR elements interact convincingly with on-screen subjects. Additionally, AI-driven 3D scene understanding allows broadcasters to create adaptive, personalized content, such as interactive replays or custom viewing perspectives. These advancements reduce production costs, increase creative flexibility, and redefine audience engagement, making AR-enhanced broadcasting more accessible and compelling across news, sports, and entertainment.

V. Conclusion and Future Work

Through the adoption of simulation platforms, such as CARLA, this study introduces the feasibility of constructing a high-fidelity dataset that encapsulate real-world complexities, including variable lighting conditions, partial occlusions, and non-static camera viewpoints. The empirical findings indicate that, while synthetic datasets may introduce challenges associated with domain adaptation, they serve as a robust framework for enhancing model generalization and performance in real-world deployment scenarios. Given the inherently time-consuming and costly nature of manual data annotation, the proposed work aims to address this limitation in the domain of motorsports scene understanding. Empirical results on real-world data demonstrate the effectiveness of the proposed dataset in minimizing the reliance on extensive labeled datasets, thereby offering a robust foundation for further analysis, and a structured way to address 3D scene reconstruction in broadcast media images.

Future research should prioritize the refinement of exclusion criteria, the development of advanced domain adaptation strategies, and the integration of physics-based simulations to further mitigate the domain gap between synthetic and real-world data. Ultimately, continued innovation in synthetic data generation methodologies will be instrumental in fostering the development of more reliable, scalable, and adaptable AI-driven vision systems for motorsports analytics and beyond.

Future work will indeed focus on advancing the end-to-end synthetic data generation pipeline, with the objective of increasing the fidelity, diversity, and domain-relevance of the generated data. Enhancements in procedural generation, domain randomization, and photorealistic rendering could significantly improve model generalization and robustness, particularly in scenarios where annotated real-world data is limited or biased. Another potential extension involves the integration of additional semantic classes, specifically targeting vehicle livery recognition. Incorporating livery as a distinct detection class would enable the system to differentiate between visually similar vehicle instances based on team or sponsor-specific visual attributes. This capability could help mitigate the inherent class imbalance and representation bias present in existing real-world datasets, thereby improving fairness and reliability in downstream perception tasks, and leading to models even more suitable for broadcasting purposes. Furthermore, a re-examination of the CARLA simulation framework presents valuable opportunities for domain-specific augmentation. By introducing racing-oriented dynamics – such as high-speed maneuvers, competitive interactions, and tactical behavior patterns – the simulation environment can be tailored to better reflect the operational context of racing environments. In this context, incorporating (inter)action recognition becomes particularly salient. Beyond static object detection, the ability to model and infer temporal and relational dynamics among agents (e.g., overtaking, blocking, cooperative maneuvers) can facilitate higher-level scene understanding and event prediction. This shift from instance-level perception to spatiotemporal reasoning has the potential to significantly enhance the decision-making capabilities of those agents operating in competitive, high-speed environments.

The generated synthetic dataset is made publicly accessible to foster further research in this field and is available for download [36].

REFERENCES

[1] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, "Computer vision for sports: Current applications and research topics", *Computer Vision and Image Understanding*, vol. 159, pp. 3–18, 2017, Computer Vision in Sports, ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2017.04.011.

[2] B. T. Naik, M. F. Hashmi, and N. D. Bokde, "A comprehensive review of computer vision in sports: Open issues, future trends and research directions", *Applied Sciences*, vol. 12, no. 9, p. 4429, 2022, ISSN: 2076-3417. DOI: 10.3390/app12094429.

[3] K. Host and M. Ivašić-Kos, "A comprehensive review of computer vision in sports: Open issues, future trends and research directions", *Heliyon*, vol. 8, no. 6, 2022, ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2022.e09633.

[4] T. Mendes-Neves, L. Meireles, and J. Mendes-Moreira, "A survey of advanced computer vision techniques for sports", *arXiv e-prints*, arXiv:2301.07583, arXiv:2301.07583, Jan. 2023. DOI: 10.48550/arXiv.2301.07583. arXiv: 2301.07583 [cs.CV].

[5] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Springer Cham, 2022. DOI: 10.1007/978-3-030-75178-4.

[6] J. Cerqueira Fernandes and B. Kenwright, "Identifying and extracting football features from real-world media sources using only synthetic training data", *arXiv e-prints*, arXiv:2209.13254, arXiv:2209.13254, Sep. 2022. DOI: 10.48550/arXiv.2209.13254. arXiv: 2209.13254 [cs.AI].

[7] G. Paulin and M. Ivasic-Kos, "Review and analysis of synthetic dataset generation methods and techniques for application in computer vision", *Artif. Intell. Rev.*, vol. 56, no. 9, pp. 9221–9265, Jan. 2023, ISSN: 0269-2821. DOI: 10.1007/s10462-022-10358-3.

[8] D. Bhargavi, E. Pelaez Coyotl, and S. Gholami, "Knock, knock. who's there? – identifying football player jersey numbers with synthetic data", *arXiv e-prints*, arXiv:2203.00734, arXiv:2203.00734, Sep. 2022. DOI: 10.48550/arXiv.2203.00734. arXiv: 2203.00734 [cs.AI].

[9] K. Duan *et al.*, "Centernet: Keypoint triplets for object detection", in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6568–6577. DOI: 10.1109/ICCV.2019.00667.

[10] H.-S. Fang *et al.*, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, Jun. 2023, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2022.3222784.

[11] H. Qin, C. Yeung, R. Umemoto, and K. Fujii, "Soccersynth-detection: A synthetic dataset for soccer player detection", *arXiv e-prints*, arXiv:2501.09281, arXiv:2501.09281, Jan. 2025. DOI: 10.48550/arXiv.2501.09281. arXiv: 2501.09281 [cs.CV].

[12] A. Cioppa *et al.*, "Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos", in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 3490–3501. DOI: 10.1109/CVPRW56347.2022.00393.

[13] Y. Cui *et al.*, " SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes ", in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 9887–9897. DOI: 10.1109/ICCV51070.2023.00910.

[14] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context", in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1.

[15] A. Boiarov, D. Bleklov, P. Bredikhin, N. Koritsky, and S. Ulasen, " RaceLens: A Machine Intelligence-Based Application for Racing Photo Analysis ", in *2023 IEEE 28th Pacific Rim International Symposium on Dependable Computing (PRDC)*, Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 352–355. DOI: 10.1109/PRDC59308.2023.00057.

[16] E. Hoffer and N. Ailon, "Deep metric learning using triplet network", in *Similarity-Based Pattern Recognition*, A. Feragen, M. Pelillo, and M. Loog, Eds., Cham: Springer International Publishing, 2015, pp. 84–92.

[17] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks", in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, Sep. 2019, pp. 6105–6114.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[19] J. Tyo, Y. Chung, M. Olarinre, and Z. C. Lipton, "Reading Between the Mud: A Challenging Motorcycle Racer Number Dataset", *arXiv e-prints*, arXiv:2311.09256, arXiv:2311.09256, Nov. 2023. DOI: 10.48550/arXiv.2311.09256. arXiv: 2311.09256 [cs.CV].

[20] I. Krylov, S. Nosov, and V. Sovrasov, "Open images v5 text annotation and yet another mask text spotter", in *Proceedings of The 13th Asian Conference on Machine Learning*, V. N. Balasubramanian and I. Tsang, Eds., ser. Proceedings of Machine Learning Research, vol. 157, PMLR, 17–19 Nov 2021, pp. 379–389.

[21] M. Huang *et al.*, "Swintextspotter: Scene text spotting via better synergy between text detection and text recognition", in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4583–4593. DOI: 10.1109/CVPR52688.2022.00455.

[22] J. Tyo, M. Olarinre, Y. Chung, and Z. C. Lipton, "Beyond the Mud: Datasets and Benchmarks for Computer Vision in Off-Road Racing", *arXiv e-prints*, arXiv:2402.08025, arXiv:2402.08025, Feb. 2024. DOI: 10.48550/arXiv.2402.08025. arXiv: 2402.08025 [cs.CV].

[23] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification", in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3701–3711. DOI: 10.1109/ICCV.2019.00380.

[24] J. Tyo and Z. C. Lipton, "Contrastive Multiple Instance Learning for Weakly Supervised Person ReID", *arXiv e-prints*, arXiv:2402.07685, arXiv:2402.07685, Feb. 2024. DOI: 10.48550/arXiv.2402.07685. arXiv: 2402.07685 [cs.CV].

[25] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator", in *Proceedings of the 1st Annual Conference on Robot Learning*, S. Levine, V. Vanhoucke, and K. Goldberg, Eds., ser. Proceedings of Machine Learning Research, vol. 78, PMLR, 13–15 Nov 2017, pp. 1–16.

[26] S. Malik, M. A. Khan, and H. El-Sayed, "Carla: Car learning to act — an inside out", *Procedia Computer Science*, vol. 198, pp. 742–749, 2022, 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.12.316.

[27] X. X. Lu, "A review of solutions for perspective-n-point problem in camera pose estimation", *Journal of Physics: Conference Series*, vol. 1087, no. 5, p. 052 009, Sep. 2018. DOI: 10.1088/1742-6596/1087/5/052009.

[28] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets", *ACM Trans. Graph.*, vol. 26, no. 3, 24–es, Jul. 2007, ISSN: 0730-0301. DOI: 10.1145/1276377.1276407.

[29] R. Mehra, P. Tripathi, A. Sheffer, and N. J. Mitra, "Visibility of noisy point cloud data", *Computers & Graphics*, vol. 34, no. 3, pp. 219–230, 2010, Shape Modelling International (SMI)

Conference 2010, ISSN: 0097-8493. DOI: https://doi.org/10.1016/j.cag.2010.03.002.

[30] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments", *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022, The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.01.135.

[31] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas", *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, 2023, ISSN: 2504-4990. DOI: 10.3390/make5040083.

[32] X. Bai *et al.*, "Bridging the domain gap between synthetic and real-world data for autonomous driving", *ACM J. Auton. Transport. Syst.*, vol. 1, no. 2, Apr. 2024. DOI: 10.1145/3633463.

[33] M. S. Werda *et al.*, "Towards minimizing domain gap when using synthetic data in automotive vision control applications", *IFAC-PapersOnLine*, vol. 58, no. 19, pp. 522–527, 2024, 18th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2024, ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2024.09.265.

[34] K. Singh, T. Navaratnam, J. Holmer, S. Schaub-Meyer, and S. Roth, "Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2024, pp. 2505–2515.

[35] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: A hands-on survey", *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 12, pp. 2633–2651, Dec. 2016. DOI: 10.1109/TVCG.2015.2513408.

[36] L. F. Rossi, *A Novel Synthetic Dataset for Broadcast Motorsports Scene Understanding*, version V1, 2025. DOI: 10.7910/DVN/DHX380.