R3ST: A Synthetic Dataset with Real Trajectories for Urban Traffic Analysis

Simone Teglia[■], Claudia Melis Tonti[■], Francesco Pro[■], Leonardo Russo[■],
Andrea Alfarano[▲], Matteo Pentassuglia[★], Irene Amerini[■]
[■]Sapienza University of Rome, [▲] INSAIT, Sofia University, [★]EURECOM, Biot, France
{*teglia, melistonti, pro, amerini*}*@diag.uniroma1.it*, *andrea.alfarano@insait.ai*,
russo.2015563@studenti.uniroma1.it, *matteo.pentassuglia@eurecom.fr*

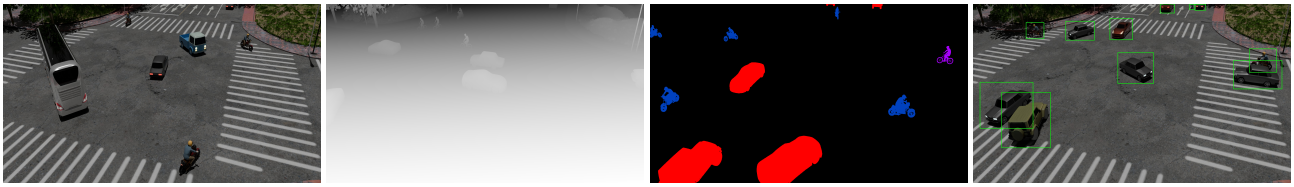

Figure 1. The R3ST Dataset. It provides photorealistic synthetic images (first), depth maps (second), instance segmentation masks (third), and object detection bounding boxes (fourth), enabling diverse computer vision tasks. Each task is presented on a different frame.

Abstract

Datasets are essential to train and evaluate computer vision models used for traffic analysis and to enhance road safety. Existing real datasets fit real-world scenarios, capturing authentic road object behaviors, however, they typically lack precise ground-truth annotations. In contrast, synthetic datasets play a crucial role, allowing for the annotation of a large number of frames without additional costs or extra time. However, a general drawback of synthetic datasets is the lack of realistic vehicle motion, since trajectories are generated using AI models or rule-based systems. In this work, we introduce ***R3ST*** *(Realistic 3D Synthetic Trajectories), a synthetic dataset that overcomes this limitation by generating a synthetic 3D environment and integrating real-world trajectories derived from SinD, a bird's-eye-view dataset recorded from drone footage. The proposed dataset closes the gap between synthetic data and realistic trajectories, advancing the research in trajectory forecasting of road vehicles, offering both accurate multimodal ground-truth annotations and authentic human-driven vehicle trajectories. We publicly release our dataset here.* [1]

1. Introduction

Understanding and studying complex urban traffic scenarios is vital for traffic analysis, urban planning, and enhancing road safety. In particular, urban intersections poses significant challenges for Traffic Monitoring Systems (TMS) due to their complex layouts and dynamic traffic conditions, as they are the most accident-prone locations in urban environments. A significant hurdle in achieving this safety is the need for a comprehensive understanding of complex traffic environments and the precise prediction of other traffic participants movements. The advancement of predicting the motion of road agents is therefore vital for developing advanced traffic monitoring systems. The analysis of traffic dynamics requires extensive datasets that effectively integrate the behavior of actual traffic participants with comprehensive annotations. Previous works like [13, 15, 16, 20] either failed to address the inherently sophisticated nature of intersections, proposing datasets that contain real trajectories but lacking the necessary realism of an intersection scenario, or presented HD-Maps based datasets, that are not always suitable for intelligent TMS due to the absence of such precise data in real time. The scarcity of a comprehensive dataset reflecting real road agent behavior poses serious challenges in creating precise and universally applicable motion prediction models, thereby constraining the efficacy of intelligent traffic monitoring in practical situations. The behavior of drivers can noticeably deviate from the expected patterns due to the intrinsic unpredictability of their intentions, bad habits, poor road design, or high traffic levels. Those factors make simulation-based traffic models insufficient to completely capture the full spectrum of human responses to varying driving conditions, despite the considerable advances of simulation environments. To

[1] https://r3st-website.vercel.app/

address these limitations and bridge the gap between realism and precise ground-truth annotations, we introduce R3ST (Realistic 3D Synthetic Trajectories), a novel large-scale synthetic dataset that uniquely integrates real, human-driven trajectories derived from actual aerial drone footage (the SinD dataset [28]) into a constructed digital twin simulation. By leveraging real-world vehicle trajectories, R3ST captures authentic traffic behaviors while providing comprehensive multimodal annotations, such as instance segmentation, depth maps, and bounding boxes, facilitated by photorealistic rendering techniques and advanced annotation tools like Vision Blender [4].

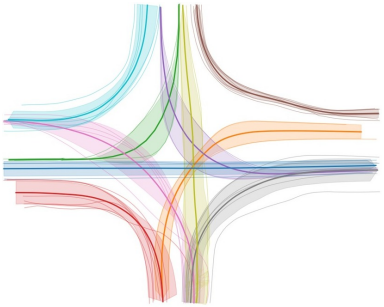

Figure 2. Visualization of clustered vehicle trajectories in a R3ST crossroad. Each colored trajectory represents a cluster of similar paths, while the shaded regions indicate variance within each cluster.

Specifically, R3ST features diverse urban intersections modelled and rendered within realistic environments created with Blender [6]. Unlike conventional synthetic datasets relying solely on rule-based or AI-generated vehicle motion, R3ST's vehicles follow real-world annotated trajectories, reflecting genuine human decision-making dynamics. Furthermore, multiple camera angles, realistic sensor parameters, and high-quality rendering settings ensure that R3ST closely mimics actual road-camera images, significantly enhancing the dataset's applicability and value for diverse computer vision tasks. Additionally, we leverage R3ST to evaluate pre-trained deep-learning models focusing on critical tasks like vehicle detection, instance segmentation, and monocular depth estimation. Through the innovative fusion of realistic human-driven trajectories and precise synthetic data generation, R3ST represents an advancement over previous datasets, offering researchers optimal capabilities for evaluating and training autonomous driving models, trajectory forecasting algorithms, and traffic analysis tools. The dataset presented in this paper offers critical foundations for future research in Intelligent Transportation Systems and contribute directly to safer, smarter urban mobility.

2. Related Work

Datasets are essential for developing and optimizing Intelligent Traffic Monitoring Systems, particularly in complex urban scenarios such as intersections. These environments pose unique challenges due to the unpredictable behavior of road users, varying lighting conditions, and the need for real-time decision-making. To address these problems, a variety of datasets have been proposed, ranging from real-world to synthetic data. In this section, we review existing datasets used for road agents' perception and motion prediction, highlighting their strengths and weaknesses. To the best of our knowledge, no existing work has explored synthetic datasets that integrate real trajectories.

Real-World Datasets Real-world datasets are collected from sensors like cameras, LiDAR, or GPS, that are placed in real-world traffic situations. Earlier works in the field of urban scene understanding like KITTI [10], CityScapes [7], SemanticKITTI [3], nuScenes [25], the Waymo Open Dataset [20] and Argoverse 2 [27] provided complex scenarios allowing tasks like object detection, semantic segmentation and trajectory prediction, but from an on-board Point Of View (POV) or relying on HD-Maps. UAVDT [9], highD [15], DroneVehicle [23], and SinD [28] instead contain images recorded by drones, giving a new point of view with respect to previous datasets, addressing object detection and object tracking. However, the creation of a real-world dataset with precise annotations of vehicle trajectories is still an ongoing challenge, due to the difficulties in acquiring large-scale, high-quality trajectory data.

Synthetic Datasets The problem of collecting real-world images and annotating them correctly, with an high expense of cost and time, has led to the creation of synthetic datasets. Synthetic datasets have the advantage of being fully controllable. They allow the generation of diverse traffic scenarios and perfectly annotate frames since the attributes of each object in the scene, like position, size, and movement, are known to the generating environment. This, without additional costs or extra time, ensures consistency and eliminates human errors in the annotation process. VIPER [18] and SYNTHIA [19] exploited game engines like Grand Theft Auto V or Unity to generate synthetic datasets. IDDA [2], SHIFT [22] and AIODrive [26] instead rely on CARLA [8], a popular simulation environment built on top of Unreal Engine 5, to create large-scale synthetic datasets for a variety of tasks, addressing the problem of the lack of various domains in a single dataset. Similarly, SynTraC [5] and Omni-IMOT [21] are among the CARLA-based datasets the most closely relate to our work, as they focus specifically on intersections. However, a significant limitation of synthetic datasets lies in the realism of vehicle movements. CARLA

Dataset	Year	Size	Real/Synthetic	POV	Depth	Instance Segmentation	Realistic Trajectories
KITTY	2012	41K frames	Real	On-Board	✗	✗	✗
nuScenes	2019	40K frames	Real	On-Board	✗	✗	✗
SinD	2019	7 hours	Real	Drone	✗	✗	✓
Argoverse2	2023	2M frames	Real	On-Board	✓	✗	✓
SYNTHIA	2016	13.4K frames	Synthetic	Street Camera	✓	✓	✗
SynTraC	2024	6 hours	Synthetic	Street Camera	✗	✗	✗
Omni-MOT	2020	14M+ frames	Synthetic	Street Camera	✗	✗	✗
SHIFT	2022	2.5M frames	Synthetic	On-Board	✗	✓	✗
R3ST	**2024**	80K+ frames	Synthetic	**Street Camera**	✓	✓	✓

Table 1. Comparison of size, nature and additional annotations of existing perception datasets. R3ST is the only dataset that leverages the advantages of synthetic data generation, combining it with realistic trajectories.

or game engines fail to recreate realistic trajectories since their movement relies on rule-based approaches that cannot model the unpredictability of human decision-making. This discrepancy poses important challenges in developing Intelligent TMS that are only trained on synthetic data, as models may struggle to generalize to real-world traffic scenarios.

3. The R3ST Dataset

Here we introduce the R3ST Dataset, a synthetic dataset that encompasses diverse urban intersections, multiple camera angles, and five vehicle types: cars, trucks, buses, motorcycles, and bicycles. This dataset has been created to leverage the benefits of synthetic data while avoiding the artificial driving behaviors typical of game engines and driving simulators like Unity of CARLA. Our dataset also provides rich annotations, including instance segmentation, depth information, and bounding boxes for object detection and object tracking. R3ST consists of photo-realistic rendering of two different intersections, each with four views, for more than 80K frames. In Table 1 we report the properties of R3ST compared with the most famous urban scene understanding datasets, highlighting that our dataset is the only one that combines the synthetic nature of the images with real trajectories.

3.1. Virtual World Generation

R3ST has been generated by rendering the virtual intersections created with Blender [6]. The motivation behind the choice of this 3D editor is the high level of freedom that it offers in creating any kind of virtual environment and the ease with which it allows the integration of external realistic trajectories. In fact, unlike typical synthetic datasets where vehicle motion is dictated by AI-driven or rule-based algorithms, R3ST incorporates real-world vehicle trajectories derived from two of the four scenarios proposed by SinD [28], a bird's-eye-view dataset with precise annotation of vehicle positions extracted from real drone footage. As visible in Figure 2, which shows the variance of clustered trajectories, this approach ensures that the motion patterns of vehicles in R3ST replicate the real-world traffic scenarios. For each trajectory of SinD, we associated the corresponding vehicle type. If the vehicle is a car we randomly selected among three different car meshes. To increase the realism further, we carefully selected materials for each mesh to achieve a visual fidelity comparable to real-world data, and also created two human models that we used as drivers for motorcycles and bicycles. We modeled two of the four scenarios presented by SinD: an intersection in Tianjin city and an intersection in Chongqing city. These virtual environments were created from scratch in Blender, utilizing external libraries such as BlenderKit [24] to import free, realistic materials, 3D vehicle models, and building models. Additionally, we employed free add-ons like Sampling Tree Generator [14], which uses a procedural algorithm to generate vegetation. This algorithm dynamically creates trees based on user-defined parameters. Furthermore, to simulate images captured from road cameras, we positioned them on light poles, setting the vertical sensor at 22mm, the field of view (FOV) to 35.3 degrees, and the F-Stop to 2.8. In the reconstructed scenes, four cameras were placed, each framing one of the four directions of traffic at the respective intersections. Vehicle animations were then rendered from the perspective of each camera.

Multimodal Annotations To enhance the usability of R3ST we leveraged Vision Blender [4] to compute additional multimodal annotations that can be used in a range of computer-vision applications. These per-frame annotations add information that goes beyond the scope of trajectory forecasting and are useful for different tasks like instance segmentation and monocular depth estimation (MDE). Additionally, we directly derive the 3D bounding box of each object in the scene from the Blender World Environment and project them on the image plane to get the 2D bounding boxes. We organize these annotations in YOLO format, where every object is encoded by its class label, the normal-

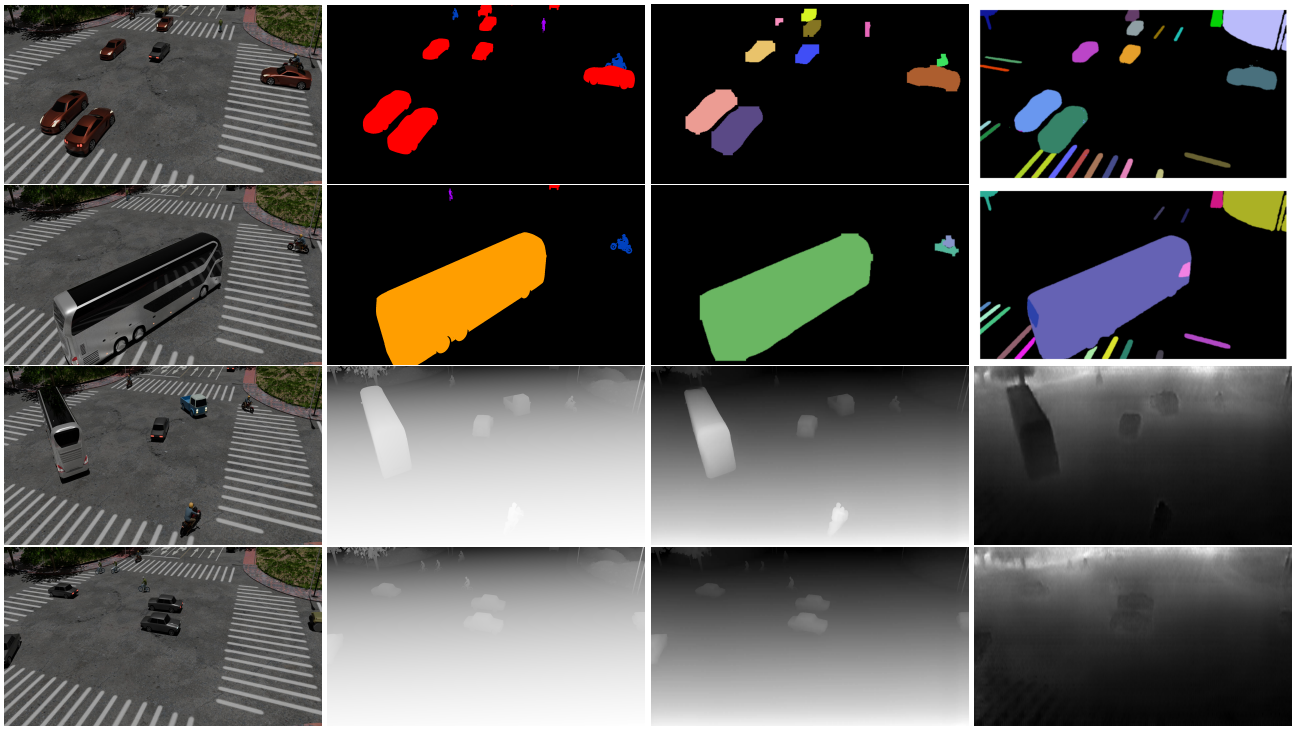

Figure 3. Qualitative results for instance segmentation (top two rows) and monocular depth estimation (bottom two rows). The first column shows RGB frames, the second column contains ground truth annotations, while the third and fourth columns present model predictions. Instance segmentation results are obtained using YOLO-Seg [12] and SAM2 [17] online demo, while monocular depth estimation is performed with AnyDepth [11] and Pixelformer Large [1] pre-trained on KITTI [10].

ized center coordinates (cx, cy), along with the normalized width and height of the bounding box.

4. Experiments

We leverage multimodal annotations to demonstrate the applicability of real world tasks to our dataset. We aim to assess how pre-trained models perform on R3ST in tasks like Object Classification, for which we show the results in Table 2, and Monocular Depth Estimation and Instance Segmentation, for which we present results in Figure 3.

Class	mAP@50	mAP@50-95
Car	0.995	0.986
Van	0.978	0.925
Motorcycle	0.990	0.972
Bicycle	0.994	0.975
Overall	0.989	0.965

Table 2. Object detection performance of YOLO11-large on the R3ST testset, composed of 900 images.

To address object detection, we tested YOLO11-large directly on R3ST, but its performance was uneven, with good precision for cars but low on motorcycles, probably due to the poor quality of the mesh used. To mitigate this, we fine-tuned YOLO11-large on R3ST for 15 epochs to show how a model trained on real images can generalize fast on our dataset. In Table 2 we report the resulting mAP@50 and mAP@50-95.

5. Conclusion and Future Works

In this paper, we introduced R3ST, the first synthetic dataset that integrates real-world vehicles trajectories, bridging the gap between realism and the flexibility of synthetic environments. We believe our dataset is valuable especially for trajectory forecasting, as well as other urban scene understanding tasks. Future works will focus on expanding the dataset with more scenes and increasing the realism of them by including different type of models for each vehicle type, as well as more challenging light conditions, to better address the domain shift problem.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention, 2022. 4

[2] Emanuele Alberti, Antonio Tavera, Carlo Masone, and Bar-

bara Caputo. Idda: A large-scale multi-domain dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5526–5533, 2020. 2

[3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2

[4] João Cartucho, Samyakh Tukra, Yunpeng Li, Daniel S. Elson, and Stamatia Giannarou. Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–8, 2020. 2, 3

[5] Tiejin Chen, Prithvi Shirke, Bharatesh Chakravarthi, Arpitsinh Vaghela, Longchao Da, Duo Lu, Yezhou Yang, and Hua Wei. Syntrac: A synthetic dataset for traffic signal control from traffic monitoring cameras. *arXiv preprint arXiv:2408.09588*, 2024. 2

[6] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 3

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator, 2017. 2

[9] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 2

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2, 4

[11] Xiankang He, Dongyan Guo, Hongji Li, Ruibo Li, Ying Cui, and Chi Zhang. Distill any depth: Distillation creates a stronger monocular depth estimator, 2025. 4

[12] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 4

[13] Vinit Katariya, Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. A pov-based highway vehicle trajectory dataset and prediction architecture. *arXiv preprint arXiv:2303.06202*, 2023. 1

[14] Lukas Kegler. Sampling tree generator - blender addon, 2017. 3

[15] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2118–2125. IEEE, 2018. 1, 2

[16] Tobias Moers, Lennart Vater, Robert Krajewski, Julian Bock, Adrian Zlocki, and Lutz Eckstein. The exid dataset: A real-world trajectory dataset of highly interactive highway scenarios in germany. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 958–964, 2022. 1

[17] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 4

[18] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks, 2017. 2

[19] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[20] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 2

[21] ShiJie Sun, Naveed Akhtar, XiangYu Song, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Simultaneous detection and tracking with motion modelling for multiple object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 626–643. Springer, 2020. 2

[22] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 2

[23] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. 2

[24] Jan Tichý. Blenderkit - online asset library for blender, 2017. 3

[25] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection, 2020. 2

[26] Xinshuo Weng, Yunze Man, Dazhi Cheng, Jinhyung Park, Matthew O'Toole, and Kris Kitani. All-In-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. *arXiv*, 2020. 2

[27] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting, 2023. 2

[28] Yanchao Xu, Wenbo Shao, Jun Li, Kai Yang, Weida Wang, Hua Huang, Chen Lv, and Hong Wang. Sind: A drone

dataset at signalized intersection in china. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2471–2478. IEEE, 2022. 2, 3