

Automated data labelling for pedestrian crossing and not-crossing actions using 3D LiDAR and RGB data

Kosmas Tsiakas^{1,2*}, Dimitrios Alexiou^{1*}, Dimitrios Giakoumis¹, Antonios Gasteratos², Dimitrios Tzovaras¹

Abstract—The existence of large-scale datasets for various autonomous driving tasks has created an increasing need for more automated annotation processes. Especially for safety critical tasks related to vehicle-pedestrian interaction, detailed and time-consuming human-made annotation is required, in order to assure accurate perception throughout any type of operating environment and for challenging conditions. In this paper, we present an automated method for the annotation of actions of humans crossing or not crossing the road. Firstly, we utilize a highly-accurate 3D multi-object tracking pipeline that combines RGB images and LiDAR data to extract the velocity and direction of movement of each pedestrian in the surrounding environment. A drivable area extraction neural network is then utilized to segment the traversable area around the vehicle. The correlation between the two above-mentioned components in the 3D space provides an accurate indication, regarding the pedestrian crossing or not-crossing the road ahead of the vehicle. Our method is validated using a custom-made multimodal dataset with an autonomous vehicle in various scenarios of a semi-structured area. The auto-generated annotations are compared directly with the human-made labels of multiple annotators and showcase the effectiveness of our method to provide an accurate indication about the human crossing the road action.

I. INTRODUCTION

The development of smart and autonomous vehicles has met great advancements over recent years, however the safety of all traffic participants cannot be guaranteed throughout the diverse traffic conditions in challenging operating environments. Various datasets have been introduced to support the progression of AI algorithms for perception [1], safety [2], path planning [3] and any other core autonomous driving task, each one containing different types of data sources [4] [5] [6]. Towards the increasing need for safe operation within urban environments, special focus has been given to the awareness about the surrounding traffic participants, the so-called vulnerable road users, through related datasets that contain information about their crossing intention [7], crossing action [8] and in general the interaction between vehicles and pedestrians in public areas [9].

A common issue of all the above mentioned datasets is the lack of standardization in terms of data format, data sources and the annotation process followed [10]. Although the quality and variety of autonomous driving datasets have significantly increased in recent years, most datasets are designed for specific scenarios or contain certain task-specific labels [11]. Con-

sequently, these datasets cannot be used by further research or applications without the inclusion of necessary additional data sources or annotations. For example, datasets that are related to pedestrian crossing intention contain solely RGB images and demonstrate critical vehicle-pedestrian interaction scenarios. Even though such data could be valuable for other tasks, e.g. multiple object tracking or trajectory prediction, the absence of LiDAR data does not permit its further re-use and development.

Recent initiatives on the mobility dataspace [12] [13] [14] focus on solving some of the above-mentioned issues through the direct exchange of data between multiple data providers and data users. The assurance of consistent data quality and standardization across different domains and sectors is highly needed, while the usage of a common data and annotation format is vital for seamless data exchange processes. However, further research is still required on the data integration pipelines, including the automated data labelling and harmonization.

The level of detail in the annotation process of datasets is highly critical to the advancement of learning-based autonomous driving tasks, especially for safety-critical tasks, such as the pedestrian crossing action and crossing intention recognition. Such procedures require extensive human labour, even with the advancement of annotation interfaces [11], while this process might sometimes be considered ambiguous, due to the biased human understanding, which is affected by familiarization with the rules of the road, signs, road delineation, and other factors, as well as the cultural background and geographical region.

In our work, we focus on the automated pedestrian "crossing" and "not-crossing" label generation for autonomous driving datasets that contain both RGB images and LiDAR data. Particularly, we combine a 3D multi-object tracking pipeline with a drivable area extraction neural network, in order to generate an accurate estimate regarding the pedestrian crossing the road action. The analysis of the relation between the pedestrians' movement and the drivable region, under a rule-based approach, is capable to provide a precise indication regarding their crossing action in every type of environment, regardless of the existence of traffic signs or marks.

The remainder of this paper is structured as follows: Section II presents a state-of-the-art analysis for datasets of pedestrians crossing and automated data annotation methods. Section III describes each component of the proposed method in detail, whereas in Section IV we evaluate the efficiency of our method. Finally, Section V concludes the paper and outlines potential directions of our research.

*Denotes equal contribution to this work

¹Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece. Email: {dalexiou, ktsiakas, dgiakoum, dimitrios.tzovaras}@iti.gr

²Department of Production and Management Engineering, Democritus University of Thrace, Xanthi, Greece. Email: agaster@pme.duth.gr

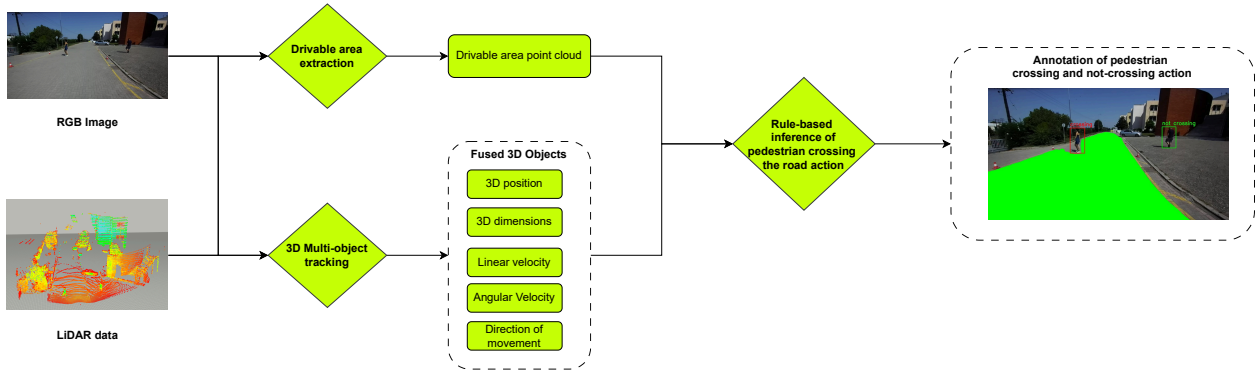


Fig. 1. The architecture of the proposed automated annotation method.

II. RELATED WORK

Recently, semi-automated and automated annotation procedures have emerged in the domain of autonomous driving datasets. Specifically, DG-Labeler [15] constitutes a semi-automated annotation procedure that utilizes depth information to infer spatial relations of instances and compute the corresponding masks for the task of multi-object tracking and segmentation. In the context of 3D object automatic annotation, MTrans [16] is a multimodal (RGB, 3D LiDAR) transformer-based approach that segments the foreground/background, densifies LiDAR point clouds, and regresses 3D boxes simultaneously from weak 2D bounding boxes. Consequently, it diminishes the significant manual effort for 3D bounding boxes annotation in LiDAR scans. Continuing with both 2D and 3D automated object detection and segmentation annotation for vision and point cloud data, OpenAnnotate3D [17] and OpenAnnotate2 [18] leverage Large Language Models (LLMs) to perform automatic annotation in multimodal data utilizing natural language prompts. Additionally, AIDE [19] constitutes a framework that not only conducts auto-labeling for object detection in images derived from autonomous driving, but also curate the corresponding data based on vision-language and large language models. The automated extraction of information about the position and locomotion of pedestrians in video streams is described in [20], towards the evaluation of pedestrian models. Regarding automation of high-level cognitive elements in autonomous driving, the authors in [21] present an offline automatic annotation process for lane markings using LiDAR and odometry data. Moreover, TADAP [22] establishes an automatic annotation process for drivable area in images under winter driving conditions.

Nevertheless, most of the works regarding automatic annotation in autonomous driving data is focused on low-level cognitive elements, i.e. 2D and 3D objects. Conversely, in our work, we propose a strategy to facilitate automation in more sophisticated perceptual elements in autonomous driving data, namely pedestrian crossing action detection, easing the generation of relevant data based on the vast amount of classical autonomous driving datasets.

The topic of pedestrian crossing action detection from the

perspective of autonomous vehicles has been widely studied by recent literature. JAAD [23] [24] is the first dataset that has been generated to study the behavior of traffic participants, mostly in urban areas. It combines both behavioral and contextual information in the point of road crossing. In the same context, PIE [25] introduced a more extensive dataset with the necessary annotations for traffic objects, pedestrians' intentions, actions and characteristics, along with ego-vehicle information. Moreover, TITAN [26] enhance 2D bounding boxes annotations of scene agents with an extensive list of pre-defined actions. Additionally, STIP [8] provides pedestrians' 2D bounding boxes along with the corresponding "crossing" or "not-crossing" the road action. Similarly, PSI [7] focuses on the crossing intention, while taking into account the temporal-dynamic state of the pedestrians. All of the above-mentioned datasets have been annotated manually through a common approach. Apart from lower-level visual annotations for traffic objects and pedestrians in image data, higher-level cognitive annotations for crossing intentions required extensive human experiments. For this process, plenty of annotators were assigned to watch multiple short-duration videos and respond to specific questions regarding the intention of pedestrians to cross the road. This process requires the occupation of multiple humans, while the results are affected by the characteristics of the annotators, e.g. their cultural background, driving skills, etc. As a result, the generation of such datasets and the evolution of pedestrian intention prediction is evolving in a slow pace.

Despite significant advances in pedestrian action and crossing intention detection, most existing methods predominantly rely on deep learning approaches trained on specific datasets with labor-intensive annotations. This reliance limits their potential generalizability across multiple autonomous driving datasets, weakening their applicability as unsupervised automatic labelling procedures. Furthermore, these methods typically utilize only RGB data, which contrasts with the emerging trend in autonomous vehicle datasets and methodologies that leverage multimodal data. In contrast, our approach employs deep neural networks for the detection of low-level perceptual elements in autonomous driving, which have demonstrated

robust generalization performance across a wide range of autonomous driving datasets.

III. PROPOSED METHOD

The overall architecture of the proposed method is depicted in Fig. 1. The inputs originate from pedestrians' tracking in 3D space and the extraction of the drivable region. The rest of the pipeline contains the rule-based inference whether the pedestrians are crossing or not the road and the corresponding annotation. Each component is described below in detail.

A. Pedestrian tracking pipeline

Tracking of pedestrians refers to the assignment of unique identifiers for each tracked object in 3D space between consecutive frames. The accuracy of this component is crucial for the overall efficient operation of our method. To this end, we propose a simple, yet effective, architecture which comprises of state-of-the-art methods for LiDAR and image-based object detection and tracking, which is depicted in Fig. 2. Each object O_i is described by a 3D centroid $C_i(x, y, z)$ and a 3D bounding box $B_i(l, w, h)$.

We use YOLOv8 [27] to perform pedestrian detection on image data. In order to convert the image pixels that correspond to a specific object into 3D space, we utilize the registered 3D LiDAR data that have common field of view with the image. Given the intrinsic matrix of the camera and the extrinsic calibration matrix between the two sensors, each point $P_i(x, y, z)$ from the input point cloud is projected into the image space $p(u, v)$ using the following equation:

$$p(u, v) = K_{intrinsic} \cdot [R|t]_{extrinsic} \cdot P(x, y, z) \quad (1)$$

This allows the generation of distinct sets of points, each one belonging to a different pedestrian, according to the values of the corresponding image mask. A geometrical clustering approach is then used in order to filter these points, remove outliers and calculate the required 3D centroid and the 3D bounding box of each object O_{2D_i} . For the rest of the paper, these object will be referred to as 2D detections.

We also perform pedestrian detection directly on 3D data using PartA2-Net [28]. This module operates on raw point cloud data and extracts all detected pedestrians, O_{3D_i} , along with the corresponding 3D centroid and 3D bounding box. For the rest of the paper, these object will be referred to as 3D detections.

The outcomes of the two above-mentioned detectors undergo a fusion process, in order to reject duplicates in the shared field of view area and ensure the most accurate representation for each pedestrian. In general, we prioritize PartA2-Net over YOLOv8 detections, due to the higher accuracy of 3D detectors irrespectively of the lightning conditions. The fusion process is relatively simple and consists of the following operation in the 3D space.

Initially, we remove duplicates between 2D and 3D detections that refer to the same object. Specifically, when $C_{2D_i}(x, y, z) \in B_{3D_j}(l, w, h)$, O_{2D_i} is removed and O_{3D_j} is appended to a new set of objects \mathcal{O} . The duplicate check is

finally performed for all objects of \mathcal{O} , so as to ensure that all pedestrians are represented by a single object and there are no overlaps, due to possible occlusions.

The subsequent step involves the tracking of the fused objects, aiming to calculate their velocity and direction of movement in 3D space. An online tracking method [29] based on Kalman filter is used for this task. The main advantage of this specific method is that it is capable of accurately tracking objects in 3D space without any prior training or extensive fine-tune of the method. For each object O_i , we obtain the linear and angular velocity $(\vec{v}_i, \vec{\omega}_i)$, as well as the direction of movement \vec{d}_i .

B. Extraction of drivable area

The extraction of the drivable area is performed initially in the image space. Following our prior work [30], we employ a multi-task network, YOLOP [31], to generate a binary mask for the drivable area. This semantic information is then converted into 3D through the utilization of the LiDAR data. Specifically, using the same transformation from Eq. 1, we generate a set of 3D points \mathcal{R} , which correspond to the drivable region in 3D space. This format allows its direct comparison between the drivable area and the tracked objects.

C. Rule-based inference of pedestrian crossing the road action

This component combines the information from the tracked objects and the drivable area, in order to automatically suggest an additional label for each pedestrian, regarding their action of "crossing" or "not crossing" the road. Following the work from [8], we consider that a pedestrian is crossing the road when their absolute value of velocity is positive and their position is within the road. It is important to note that when a pedestrian enters the road region, the drivable area extraction module calculates the drivable area surrounding the pedestrian, as illustrated in Fig. 3.

Consequently, in order to infer the crossing action we perform a geometrical check regarding the pedestrian's position $C_i(x, y, z)$ compared to the drivable area \mathcal{R} in 3D space. Specifically, we deposit the drivable region point cloud in a K-D tree structure \mathcal{T} and then we conduct a radius search of $0.5m$ for the pedestrians' centroid to determine its nearest neighbors, denoted as n_i , for the pedestrian O_i .

$$n_i = \{R_j \in \mathcal{R} \mid \|C_i - R_j\| \leq 0.5\} \quad (2)$$

Then we employ a heuristic threshold filtering (Eq. 3), to determine whether the pedestrian is inside the road or not.

$$n_i > 5 \quad (3)$$

For the final inference of pedestrian crossing action, we additionally check if the absolute value of linear velocity \vec{v}_i of the tracked pedestrians is positive.

$$|\vec{v}_i| > 0 \quad (4)$$

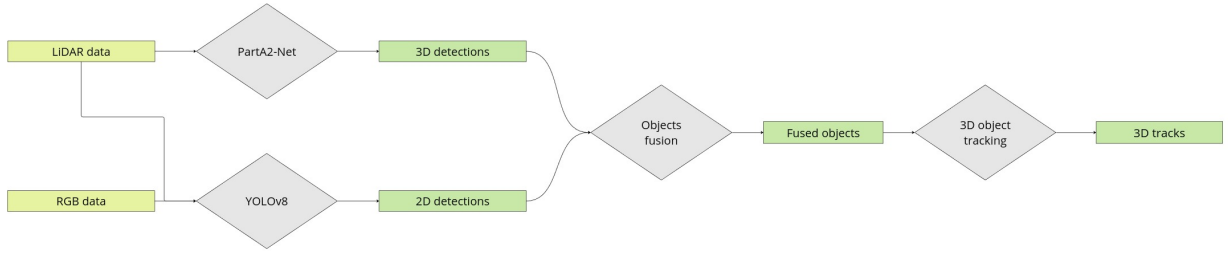


Fig. 2. The architecture of the proposed 3D multi-object tracking pipeline using LiDAR and RGB data.

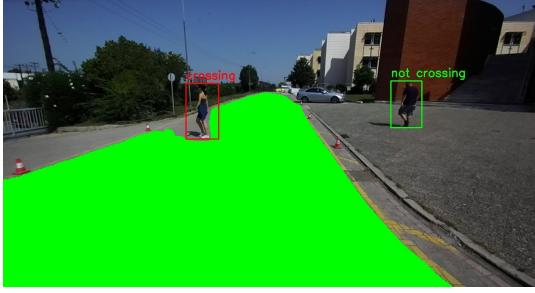


Fig. 3. Drivable area extraction in the presence of a pedestrian. The drivable area extraction module dynamically adjusts, computing the drivable area surrounding the pedestrian.

IV. EXPERIMENTS

In order to verify the effectiveness of our method, we performed a dedicated data collection using our autonomous vehicle in the semi-structured premises of CERTH. The existing pedestrian crossing datasets contained only RGB images, making it impossible to use our method to reproduce the existing labels. For this reason, it was not feasible to validate our method in such a dataset. So we constructed a dataset similar to most of autonomous driving datasets with a high resolution 3D LiDAR, namely Livox HAP, facing forward, a front-facing ZED 2 stereo camera, a GPS and an IMU. The data collection was performed in an intersection, in a large-scale area with no traffic marks and in an one-way road with the pedestrians acting randomly.

During the data collection process, pedestrians were assigned to walk in the surrounding area of the vehicle and either cross the street, walk in the non-roadside areas or remain stationary. The dataset comprises of 100 randomly sampled scenes that include both RGB images and 3D LiDAR measurements. Within these scenes, there are 93 instances of pedestrians not engaged in crossing actions and 81 instances of pedestrians engaged in crossing actions. Once the data collection was completed, we performed the data annotation, in order to manually assign a label for each pedestrian that appeared in the images as "crossing" or "not-crossing". In order to make this process accurate and follow the existing literature on this topic, we followed the approach from [7]. Specifically, 5 people were assigned the task of annotation, each one being totally neutral about the process of data

collection and the overall concept of pedestrian crossing in the datasets. An average of the assigned labels was calculated and through our validation, we aim to evaluate the accuracy of our method to automatically suggest the labels of "crossing" or "not-crossing", as closely as possible to the humans' perception.

A quantitative evaluation was initially performed to validate the accuracy of our method utilizing the mean Average Precision (mAP) metric similar to the approach described in [32]. A detection was considered as true positive if the predicted and ground truth bounding boxes share the same label and the Intersection over Union (IoU) exceeded the thresholds of 0.5 and 0.75, respectively. The results of the quantitative evaluation are presented in Table I.

TABLE I
RESULTS ON CERTH DATASET.

values in %	Overall mAP	Not-crossing mAP	Crossing mAP
$\text{IoU} \geq 50$	82.21	75.56	88.86
$\text{IoU} \geq 75$	71.19	56.06	86.32

In Fig. 4, indicative qualitative results of our methods are depicted. The top row displays the results of our automated data labelling process, distinguishing between pedestrian "crossing" and "not-crossing" actions, while the bottom row depicts the corresponding ground truth annotations. This comparison illustrates the efficacy of our labelling method in accurately identifying the pedestrian crossing action. In general, our method annotates successfully in most of the cases as it can be seen from both quantitative and qualitative results. However, slight human verification and correction is needed, as indicated by the mean average precision metrics. This step is helpful for addressing undetected and erroneously detected results, particularly in challenging conditions such as distant pedestrians and occlusions. Despite this, the required annotation time remains significantly lower compared to annotating from scratch. It is important to note that our pipeline is characterised from high modularity, as it can be further improved by the utilization of superior 2D and 3D detectors, drivable area segmentation models and object tracking modules.

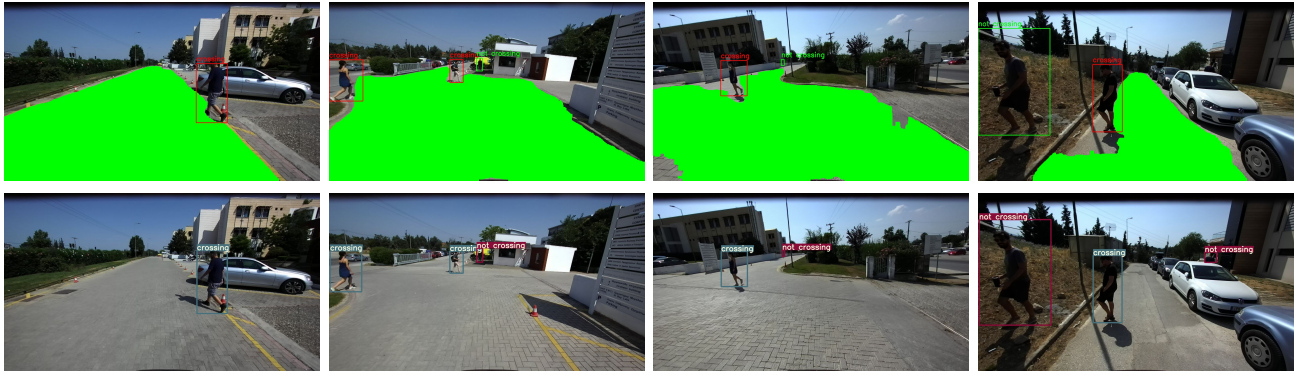


Fig. 4. Top row: qualitative results of our automated data labelling process for pedestrian "crossing" or "not-crossing". Bottom row: ground truth.

V. CONCLUSIONS

An automated method for data generation of pedestrian crossing actions was introduced in this paper. Specifically, a simple yet effective pipeline for multi-object tracking in 3D space, using LiDAR and image data, was introduced. In combination with a drivable area extraction neural network, our method can distinguish between the "crossing" and "not-crossing" pedestrians through the comparison of pedestrians' velocity, position, and their relation to the drivable area. Our method is validated on a dataset collected in the premises of CERTH, demonstrating its effectiveness in generating reliable labels similar to those human annotators would generate. Further research is required in this domain, particularly in extending this approach to additional tasks, such as predicting pedestrian intention, while further sources of data could be investigated to support such tasks.

ACKNOWLEDGMENT

This work has been supported by the EU Horizon Europe funded project "PLIADES" under the GA No: 101135988.

REFERENCES

- [1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [2] T. Wang, S. Kim, J. Wenxuan, E. Xie, C. Ge, J. Chen, Z. Li, and P. Luo, "Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5599–5606.
- [3] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.
- [4] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards deep radar perception for autonomous driving: Datasets, methods, and challenges," *Sensors*, vol. 22, no. 11, p. 4208, 2022.
- [5] I. Lahouli, E. Karakasis, R. Haelterman, Z. Chtourou, G. De Cubber, A. Gasteratos, and R. Attia, "Hot spot method for pedestrian detection using saliency maps, discrete chebyshev moments and support vector machine," *IET Image processing*, vol. 12, no. 7, pp. 1284–1291, 2018.
- [6] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.
- [7] T. Chen, T. Jing, R. Tian, Y. Chen, J. Domeyer, H. Toyoda, R. Sherony, and Z. Ding, "Psi: A pedestrian behavior dataset for socially intelligent autonomous car," *arXiv preprint arXiv:2112.02604*, 2021.
- [8] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.
- [9] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.
- [10] Y. Wang, Z. Han, Y. Xing, S. Xu, and J. Wang, "A survey on datasets for the decision making of autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, 2024.
- [11] M. Liu, E. Yurtsever, J. Fossaert, X. Zhou, W. Zimmer, Y. Cui, B. L. Zagar, and A. C. Knoll, "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.
- [12] C. S. Langdon and K. Schweichhart, "Data spaces: First applications in mobility and industry," 2022.
- [13] R. A. Deshmukh, D. Collarana, J. Gelhaar, J. Theissen-Lipp, C. Lange, B. T. Arnold, E. Curry, and S. Decker, "Challenges and opportunities for enabling the next generation of cross-domain dataspace," in *The Second International Workshop on Semantics in Dataspace, co-located with the Extended Semantic Web Conference*, 2024.
- [14] C. Doukeridis, G. M. Santipantakis, N. Koutroumanis, G. Makridis, V. Koukos, G. S. Theodoropoulos, Y. Theodoridis, D. Kyriazis, P. Kranas, D. Burgos *et al.*, "Mobispaces: An architecture for energy-efficient data spaces for mobility data," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 1487–1494.
- [15] Y. Cui, Z. Cao, Y. Xie, X. Jiang, F. Tao, Y. V. Chen, L. Li, and D. Liu, "Dg-labeler and dgl-mots datasets: Boost the autonomous driving perception," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 58–67.
- [16] C. Liu, X. Qian, B. Huang, X. Qi, E. Lam, S.-C. Tan, and N. Wong, "Multimodal transformer for automatic 3d annotation and object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 657–673.
- [17] Y. Zhou, L. Cai, X. Cheng, Z. Gan, X. Xue, and W. Ding, "Openannotate3d: Open-vocabulary auto-labeling system for multi-modal 3d data," *arXiv preprint arXiv:2310.13398*, 2023.
- [18] Y. Zhou, L. Cai, X. Cheng, Q. Zhang, X. Xue, W. Ding, and J. Pu, "Openannotate2: Multi-modal auto-annotating for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024.
- [19] M. Liang, J.-C. Su, S. Schuster, S. Garg, S. Zhao, Y. Wu, and M. Chandraker, "Aide: An automatic data engine for object detection in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 695–14 706.
- [20] E. Boukas, L. Crociani, S. Manzoni, G. Vizzari, A. Gasteratos, and G. C. Sirakoulis, "An intelligent tool for the automated evaluation of pedestrian simulation," in *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings*. Springer, 2014, pp. 136–149.
- [21] J. B. Martirena, M. N. Doncel, A. C. Vidal, O. O. Madurga, J. F. Esnal, and M. G. Romay, "Automated annotation of lane markings using lidar and odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3115–3125, 2020.

- [22] E. Alamik Kotervo, R. Ojala, A. Seppänen, and K. Tammi, "Tadap: Trajectory-aided drivable area auto-labeling with pretrained self-supervised features in winter driving conditions," *IEEE Transactions on Intelligent Vehicles*, 2024.
- [23] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 206–213.
- [24] —, "Agreeing to cross: How drivers and pedestrians communicate," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 264–269.
- [25] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6262–6271.
- [26] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 186–11 196.
- [27] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [28] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2647–2664, 2020.
- [29] H.-k. Chiu, A. Prioletti, J. Li, and J. Bohg, "Probabilistic 3d multi-object tracking for autonomous driving," *arXiv preprint arXiv:2001.05673*, 2020.
- [30] K. Tsiakas, D. Alexiou, D. Giakoumis, A. Gasteratos, and D. Tzovaras, "Leveraging multimodal sensing and topometric mapping for human-like autonomous navigation in complex environments," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7415–7421.
- [31] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, vol. 19, no. 6, pp. 550–562, 2022.
- [32] J. Cartucho, R. Ventura, and M. Veloso, "Robust object recognition through symbiotic deep learning in mobile robots," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 2336–2341.