Video Killed the HD-Map: Predicting Multi-Agent Behavior Directly From Aerial Images

Yunpeng Liu^{1,2} Vasileios Lioutas^{1,2} Jonathan Wilder Lavington^{1,2} Matthew Niedoba^{1,2} Justice Sefas^{1,2} Setareh Dabiri¹ Dylan Green^{1,2} Xiaoxuan Liang^{1,2} Berend Zwartsenberg¹ Adam Ścibior¹ Frank Wood^{1,2,3}

Abstract—The development of algorithms that learn multiagent behavioral models using human demonstrations has led to increasingly realistic simulations in the field of autonomous driving. In general, such models learn to jointly predict trajectories for all controlled agents by exploiting road context information such as drivable lanes obtained from manually annotated high-definition (HD) maps. Recent studies show that these models can greatly benefit from increasing the amount of human data available for training. However, the manual annotation of HD maps which is necessary for every new location puts a bottleneck on efficiently scaling up human traffic datasets. We propose an aerial image-based map (AIM) representation that requires minimal annotation and provides rich road context information for traffic agents like pedestrians and vehicles. We evaluate multi-agent trajectory prediction using the AIM by incorporating it into a differentiable driving simulator as an image-texture-based differentiable rendering module. Our results demonstrate competitive multi-agent trajectory prediction performance especially for pedestrians in the scene when using our AIM representation as compared to models trained with rasterized HD maps.

I. INTRODUCTION

Creating realistic simulation environments is crucial for evaluating self-driving vehicles before they can be deployed in the real world. Recent studies have emphasized the use of learned models to generate more realistic behavior for controlled agents like pedestrians and surrounding vehicles [1]– [3]. These models learn to imitate human-like behaviors in a traffic scene by utilizing a probabilistic conditional model of multi-agent trajectories in an environment. When using such approaches to construct realistic simulations, the quality of learned behavior is strongly dependent on the amount of data used for training [4], [5].

Typically, learning behavior models requires data consisting of the high-definition (HD) map for the given location and extracted agent tracks. While the latter can be extracted automatically from sensory data using modern computer vision algorithms with good accuracy [4]–[6], doing that for the former is still an open problem [7]–[9] and in practice, manual annotations are often used. Moreover, since it is important to ensure not only a large number of hours in the dataset but also a large variety of locations, manually annotating HD maps can become the most laborious part of creating a dataset. To make things worse, HD maps inevitably fail to capture important context, and increasing their detail like annotating sidewalks and crosswalks (see Fig. 2b) increases the cost of annotation. For example, the



Fig. 1: (a) An example of a simulated scene with the rasterized HD map representation compared to (b) the aerial image-based map (AIM) representation rendered using our image-texture-based differentiable rendering module. The AIM representation requires minimum annotation effort as it is obtained directly from (c) the raw drone video recording frame with agents removed (d). Orange circles in (b) highlight examples of rich road context information.

simplistic HD map scheme used in Fig. 1a does not reflect pedestrian crosswalks, sidewalks and bus lane designations.

In this study, we investigate the performance of behavioral models learned using aerial imagery instead of HD maps. Specifically, we record a dataset of human behavior in traffic scenes with a drone from a bird's-eye view, in a manner similar to [11], and extract the background aerial image by averaging the collected video frames of the location. While other background extraction techniques can be applied [12]–[14], we find this simple averaging approach is sufficient for our use case. We refer to this image as the "aerial image-based map" (AIM), emphasizing that it is both easy to obtain automatically and that it contains rich contextual information.

Learning trajectory prediction models by behavioral cloning is known to suffer from the covariate shift, where prediction quality drops drastically with simulation time, and it has been demonstrated that this issue can be ameliorated by imitating in a differentiable simulator instead [2], [3],

¹Inverted AI, ²University of British Columbia, ³Mila



Fig. 2: Examples of HD maps from public motion planning datasets for (a) Argoverse [10] and (b) Nuplan. The Nuplan map includes annotations like crosswalks and parking areas (shown in orange), which are not labeled in the Argoverse map.

[15]. We use a similar approach incorporating the AIM into a differentiable simulator by implementing a custom differentiable renderer. The renderer, illustrated in Fig. 3, uses the AIM as background and places simple rasterizations of agents and traffic lights on top of it, as shown in Fig. 1b. To evaluate the impact of using AIM, we employ a multiagent trajectory prediction model, ITRA [3] which consumes rasterized views of HD maps shown in Fig. 1a. We compare ITRA trained with AIM representation (ITRA-AIM) shown in Fig. 1b with the same model using HD map representation (ITRA-HDM) for two dominant traffic agent categories, pedestrians and vehicles. ITRA-AIM demonstrates competitive performance compared to ITRA-HDM on widely used metrics such as minimum Average Displacement Error (minADE) and minimum Final Displacement Error (minFDE) for both agent types. Notably, ITRA-AIM exhibits an even higher performance gain on pedestrian trajectory prediction.

II. BACKGROUND

In this section, we will formally define our multi-agent trajectory prediction problem and give an overview of ITRA, the multi-agent trajectory prediction model which we use to evaluate our AIM representation. We will also introduce the concept of differentiable driving simulators which are applied in many trajectory prediction models, including ITRA.

A. Multi-agent trajectory prediction

In this paper, we define the state for N agents across T time steps as s_T^N (following the notation used in [3]). For a specific agent i, its state $s_t^i = (x_t^i, y_t^i, \phi_t^i, v_t^i)$ for $t \in 1, \ldots, T$ consists of the agent's coordinates, as well as its direction and velocity relative to a stationary global reference frame. In the multi-agent trajectory prediction setting, we are interested in predicting the future joint state $s_{tobs+1:T}^{1:N}$, given t_{obs} state observations while conditioning on the road context information. This information is traditionally represented by a so-called HD map. These maps consist of road polygons, lane boundaries, lane directions, and may also include additional features such as crosswalks.

B. ITRA

We use ITRA [3] to investigate the validity of our primary claim. ITRA uses a conditional variational recurrent neural network (CVRNN) [16] model followed by a bicycle kinematic model [17] to jointly predict the next state of each agent in the scene. All interactions between the agents and the environment are encoded using differentiably rendered birdview images. These birdview representations are centered at the agent of interest and rotated to match its orientation. Each agent i at timestep t is modeled as a CVRNN with recurrent state h_t^i and latent variables $z_t^i \sim \mathcal{N}(z_t^i; 0, \mathbf{I})$. After ITRA obtains the corresponding ego-centered birdview $b_t^i = \texttt{render}(i, s_t^{1:N_t}), \text{ it produces the next action } a_t^i =$ $f_{\theta}(b_t^i, z_t^i, h_{t-1}^i)$, where $h_t^i = f_{\psi}(h_{t-1}^i, b_t^i, a_t^i)$ is generated using a recurrent neural network. The next state $s_{t+1}^i \sim$ $\mathcal{N}(s_{t+1}^i; f_{kin}(s_t^i, a_t^i), \sigma \mathbf{I})$ is produced using a kinematic bicycle model f_{kin} and the generated action a_t^i . The joint model $p(s_{1:T}^{1:N})$ factorizes as

$$\iint \prod_{t=1}^{T} \prod_{i=1}^{N} p(s_{t+1}^{i} | s_{t}^{i}, a_{t}^{i}) p_{\theta}(a_{t}^{i} | b_{t}^{i}, z_{t}^{i}, h_{t-1}^{i}) \\ p(z_{t}^{i}) dz_{1:T}^{1:N} da_{1:T}^{1:N}.$$
(1)

The model is trained jointly with a separate inference network $q_{\phi}(z_t^i|b_t^i, a_t^i, h_{t-1}^i)$ by maximizing the evidence lower bound (ELBO),

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}_{q_{\phi}(z_{t}^{i}|b_{t}^{i},a_{t}^{i},h_{t-1}^{i})} \left[\log p_{\theta}(s_{t+1}^{i}|b_{t}^{i},z_{t}^{i},h_{t-1}^{i}) \right] \\ - \mathrm{KL} \left[q_{\phi}(z_{t}^{i}|b_{t}^{i},a_{t}^{i},h_{t-1}^{i}) || p(z_{t}^{i}) \right],$$
(2)

where $p_{\theta}(s_{t+1}^{i}|b_{t}^{i}, z_{t}^{i}, h_{t-1}^{i}) = \int p(s_{t+1}^{i}|s_{t}^{i}, a_{t}^{i}) p_{\theta}(a_{t}^{i}|b_{t}^{i}, z_{t}^{i}, h_{t-1}^{i}) d_{a_{t}^{i}}.$

C. Differentiable driving simulators

Previous research has demonstrated that performing imitation learning within a differentiable simulator can help mitigate the distributional shift due to compounded error in open-loop behavior cloning methods [2], [15], [18]. These simulators typically consist of a differentiable kinematic model, which produces the next state s_{t+1}^i given the current state-action pair. Additionally, they have a differentiable renderer that generates the ego-centered bird's-eye view image that includes the road context information and other agents in the scene. One of the main advantages of using such differentiable simulators is that the loss in Eq. (2) can be directly optimized using backpropagation as the state transition $p(s_{t+1}^i | s_t^i, a_t^i)$ is fully differentiable.

III. RELATED WORK

Current methods to multi-agent modeling approach the problem by jointly predicting future trajectories using deep probabilistic models such as conditional variational auto-encoders (CVAEs) [2], [3], [19], [20], normalizing flows [21], [22] and more recently diffusion models [23]. This family of multi-agent trajectory prediction models relies heavily on obtaining road context from HD maps, which requires manual annotations of lane center and boundary

lines. To represent such HD maps, one approach is to render the semantic information of the map into a birdview image by employing different color schemes [2], [3], [15]. This image is then encoded using a convolutional neural network (CNN). Alternatively, recent work [24], [25] suggests representing road elements as a sequence of vectors that can be employed by a graph neural network (GNN) which achieves better performance than the rendering approach. However, regardless of the embedding method, an annotated map has to be obtained, which our method bypasses completely. Moreover, aerial images contain rich road context information such as left and right turn road markings and bus lanes without any annotation effort. In prior research [26], satellite imagebased maps were used as a substitute for HD semantic maps, and the resulting findings indicated that trajectory prediction exhibited worse performance in such images. We point out that in those cases, satellite map representation that contains traffic light states were not evaluated and the visual quality of the road context is limited. In contrast, our AIM representation does contain traffic light states, and the representation of agent states obtained from our differentiable rendering module has a higher visual quality. We demonstrate in Sec. V-D that these two factors can have significant impact on the prediction performance.

Previous studies have shown that integrating scene context into pedestrian trajectory prediction models improves their performance [27]. These studies [28], [29] often encode a single static bird's-eye view image of the scene using a CNN to represent scene information. The sequence of video frames can also serve as input for scene context information using graph convolutional neural networks [30]. To incorporate agent information, these methods utilize a separate network, and scene information is merged with agent information using attention [28], GNNs [30] or CNNs [29]. However, the aforementioned methods suffer from covariate shift over a long time, limiting their application for simulation purposes.

IV. PROPOSED METHOD

Our method incorporates unlabelled aerial images into a simulation environment [3] using a differentiable renderer implemented with Pytorch3D [31]. We leverage image-texture-based rendering to represent the background of the simulated scene, as depicted in Fig. 1b. By embedding this rendering module into an existing end-to-end differentiable 2D driving simulator that targets multi-agent trajectory prediction, we can produce ego-rotated and egocentric birdview images b_t^i that utilize the proposed aerial image-based map (AIM) for multi-agent trajectory prediction models like ITRA. Furthermore, it provides a representation of the road context with minimal information loss, as opposed to a rasterized birdview image from the labeled HD map shown in Fig. 1a. We introduce our image-texture-based differentiable rendering module in the following section.

A. Image-texture-based differentiable rendering.

Our image-texture-based rendering module is designed to be differentiable and efficient as it supports rendering in



Fig. 3: Image-texture-based rendering procedure.

TABLE I: Validation set prediction errors on our pedestrian dataset.

Method	$minADE_6\downarrow$	$minFDE_6\downarrow$	$MFD_6\uparrow$
ITRA-P-HDM	0.90	2.01	0.29
ITRA-P-AIM	0.74	1.56	0.51

batch mode. The rendering process is illustrated in Fig. 3. Our module takes in the processed drone image of the recording location, along with its coordinates of the four image corners $\{x_c, y_c\}_{c=0}^4$, the states of the agents $s_t^{1:N}$ at time t, and the traffic light states. We average motion stabilized drone video frames to perform a simple yet effective background extraction of the video. This background extraction process serves to eliminate cars and other agents that may be present in the drone video. To differentiate agent types, each agent type is associated with a unique color in the texture map (see Fig. 3). The image corner coordinates are in the same global reference frame as the agent coordinates to align the map with the agents during the rendering process. Using the aforementioned inputs, three distinct types of meshes are constructed, namely the background mesh, agent mesh and the traffic light mesh along with their corresponding texture maps. These meshes are subsequently combined to form a concatenated mesh M with a merged texture map, which is then fed into a differentiable renderer, to render the simulated scene. Agents are rendered as bounding boxes with an additional triangle on top of each bounding box to indicate their direction. Traffic lights are rendered as rectangular bars at the stop line. The rendered birdview can be consumed by the trajectory prediction models as a representation of the environment for the ego agent which provides information about the road context and other agents in the scene.



Fig. 4: **Column (a-c)** Ego-only trajectory prediction of 40 timesteps based on the observation of only one initial timestep for vehicle agents. We show 10 sampled trajectories in orange alongside the ground truth trajectory colored in grey for both ITRA-V-HDM and ITRA-V-AIM. Note that ITRA trained with the AIM representation generates more realistic samples by leveraging the road context, particularly in scenarios such as entering parking lots. **Column (d-e)** Examples of Multi-agent trajectory predictions on two different map representations.



Fig. 5: Column (a-c) Ego-only trajectory prediction of twelve seconds based on the observation of only one initial observation for pedestrians. We show 5 sampled trajectories in orange alongside the ground truth trajectory colored in grey for both ITRA-P-HDM and ITRA-P-AIM. Pedestrians are highlighted with orange bounding boxes in ITRA-P-AIM for visualization purposes. Column (d-e) Examples of Multi-agent trajectory predictions on two different map representations. Note that the pedestrian model trained with AIM tends to navigate within designated areas such as sidewalks and crosswalks. This behavior is facilitated by the incorporation of comprehensive road context information provided by the AIM, enabling the model to leverage the contextual cues to figure out pedestrians movement patterns.

V. EXPERIMENTS

We evaluate the effectiveness of our AIM representation with ITRA (ITRA-AIM) on a dataset comprising 5.5 hours of traffic data collected using a commercial drone in 20 locations primarily in Canada. These locations include roundabouts, signalized and unsignalized intersections, and highways, providing a diverse set of road geometries for training and evaluation. We carefully selected 11 locations out of the 20 locations for training the ITRA-AIM model on vehicles. These particular locations were chosen based on the presence of rich driving behavior. Similarly, we selected 17 locations specifically for training our model on pedestrian data, considering that pedestrians are typically absent from locations such as highways.

Our dataset comprises 300k four-second long segments sampled at 10 Hz of vehicle data. On the other hand, the

TABLE II: Validation set prediction errors on our vehicle dataset.

Metrics	ITRA-V-HDM	ITRA-V-AIM	ITRA-V-AIM-ResNet18
$minADE_6\downarrow$	0.50	0.45	0.50
$minFDE_6\downarrow$	1.04	0.93	1.10
Off-road rate↓	0.006	0.008	0.006
collision rate↓	0.012	0.012	0.012

pedestrian dataset consists of approximately 200k segments, each spanning twelve seconds and sampled at a rate of 2.5 Hz. This lower sampling rate is employed due to the comparatively slower movement of pedestrians compared to vehicles, which is consistent with other well-known pedestrian datasets [32], [33]. We reserved the final 5% of our drone recordings obtained at each location as our validation dataset, ensuring that the training data has no causal relationship to the validation data.

In our ablation studies, we demonstrate the importance of rendering traffic lights in the AIM, as well as the impact of aerial image quality on displacement errors and infraction rates of the generated samples. These findings provide insights to why prior work [26] obtained inferior results on satellite images as mentioned in Sec. III. We also test on an aerial image from Bing [34] at one of our recording locations.

A. Implementation details

To demonstrate the impact of AIM on motion prediction for pedestrian and vehicle agents, we train separate models for the two agent types (ITRA-V and ITRA-P). In addition, to compare ITRA-AIM with the original ITRA that utilizes the rasterized HD map (ITRA-HDM), we apply the same training procedure on ITRA-AIM as our baseline, training each component of the network from scratch and using the same training hyper-parameters for ITRA-AIM and ITRA-HDM. We use an identical CNN encoder for encoding AIM which consists of a 4-layer CNN model for our ITRA-AIM model but also experiment with a ResNet-18 backbone on the vehicle dataset to encode the AIM representation, because it contains more information than the rasterized HD map. While the training setup is the same for ITRA-V and ITRA-P, we apply a unicycle dynamics model [35] as f_{kin} for pedestrians instead of the bicycle model we use for vehicles.

ITRA adopts classmates forcing [20] in the training phase, which provides ground truth states to the model for all agents beside the ego agent. At test time, ITRA trained on the vehicle dataset (ITRA-V) jointly predicts the future trajectories for all vehicles while other agent types like pedestrians are replayed in the scene with ground truth trajectories. Similarly, ITRA-P predicts the motion of all pedestrians in the scene given the ground truth vehicle trajectories. We train all of our models with a random observation length between 1 to 10 timesteps to prevent overfitting on past observations and they are trained until the validation loss converges. The training time of ITRA-AIM is comparable to that of ITRA-HDM on 4 NVIDIA GeForce RTX 2080 Ti GPUs.

B. Evaluation Metrics

Common metrics for evaluating trajectory prediction models are ADE and FDE, which measure how close the sampled trajectory is to the ground truth trajectory. In the multiagent setting, ADE and FDE are averaged across all N agents. Given K trajectory prediction samples, the generated trajectory with the minimum error is selected for calculating $minADE_K$ and $minFDE_K$. As AIM does not explicitly label the drivable area, we evaluate the prediction performance of the ITRA-V-AIM model on how often it commits off-road infractions. To accomplish this, we calculate the off-road rate using the method described in [36], assuming access to a drivable surface mesh for evaluation purposes. The computed off-road rate is zero when all four corners of the vehicle are within the drivable area. Furthermore, we also measure the collision rate in a multi-agent prediction setting using the intersection over union (IOU)-based collision metric from [36]. Both the off-road rate and collision rate reported in Tabs. II to IV are averaged across the number of agents, samples, and time. Since AIM provides additional road context information like sidewalks and crosswalks, we also measure the diversity of trajectories generated for pedestrians by measuring the Maximum Final Distance (MFD) averaged over number of agents introduced in [3].

C. Experimental results

We present our results on the validation dataset for ITRA-P in Tab. I and for ITRA-V in Tab. II. In the case of ITRA-P, we jointly predict the eight-second future given the initial four-second observation. As for ITRA-V, we predict the trajectory for a time horizon of 40 timesteps (four seconds) while observing the first 10 timesteps. On the pedestrian dataset, ITRA-P-AIM outperforms ITRA-P-HDM across all evaluation metrics and achieves higher diversity. This result indicates the importance of providing scene context information for pedestrians when modeling them in traffic simulations. We showcase our predicted examples on the validation dataset in Fig. 5. ITRA-V-AIM demonstrates competitive performance in reconstructing ground truth trajectories on the validation dataset. Regarding the off-road rate, ITRA-V-AIM with a ResNet-18 backbone matches the performance of ITRA-V-HDM while maintaining similar displacement errors. Figure 4 presents validation examples comparing ITRA-V-HDM and ITRA-V-AIM on the vehicle dataset.

D. Ablation Studies

To investigate the impact of incorporating traffic light states into the AIM representation, we analyze their influence on prediction results, since previous work [26] did not evaluate representations that include traffic light states. Specifically, we select locations from our dataset where traffic light states are available and conduct a comparative evaluation of the ITRA-V-AIM model on AIMs with and without traffic light states.

Our results, presented in Table III, demonstrate a notable reduction in collision rates of over 15% and a 12% decrease in the minADE metric when traffic light states are rendered



(a) Original aerial image(b) Degraded aerial imageFig. 6: Comparison of aerial image quality.

TABLE III: Comparison between with and without traffic light rendered on locations have traffic light labels.

	ITRA-V-AIM		
Metrics	With traffic lights	Without traffic lights	
$minADE_6\downarrow$	0.37	0.42	
$minFDE_6\downarrow$	0.77	0.88	
Off-road rate↓	0.006	0.008	
Collision rate↓	0.011	0.013	

on AIMs. Additionally, we observe a similar performance improvement for our pedestrian model, ITRA-P-AIM, when traffic light states are incorporated. We attribute part of the AIM's competitive performance to our proposed imagetexture-based differentiable rendering module, which enables the integration of traffic light states within the AIM.

We also study the effect of aerial image quality on prediction results. Depending on the height at which the image is captured and camera's specifications, the image quality, in terms of resolution and noise level, can vary drastically. To simulate a deterioration in these quality factors, we first apply Gaussian blur to our original aerial images and then add Gaussian noise with a standard deviation of 0.2 to these blurred images (an example is shown in Fig. 6b). We evaluate ITRA-V-AIM on the degraded AIMs using the validation agent tracks and report the results in Tab. IV. The degraded AIMs result in increased prediction errors and nearly double the off-road infraction rate. Although the simulated degradation may not precisely emulate the various aspects of the reduction in image quality, our findings highlight the crucial role of capturing high fidelity imagery in ensuring the accuracy of multi-agent trajectory prediction through our AIM representation.

To demonstrate the flexibility of our AIM representation, we acquired an aerial image from Bing aerial imagery of the same location as Fig. 4d to construct a larger AIM. Despite having different image conditions (such as shading and lighting) compared to our original AIM, our trained ITRA-V-AIM achieves good prediction performance on this larger map with validation data segments as shown in Fig. 7.

VI. CONCLUSION

In this work, we have addressed a critical bottleneck in scaling the dataset size for behavioral models used for simulating realistic driving. Specifically, the manual annotation of high-definition maps on new locations impedes progress towards the fully automated labeling of datasets. Rather than pursuing the development of automated map labeling tools, which may introduce additional labeling noise, our

TABLE IV: The impact of aerial image quality on prediction results.

	ITRA-V-AIM		
Metrics	Original AIM	Blurred and noise-added AIM	
minADE ₆	0.45	0.49	
minFDE ₆	0.93	1.04	
Off-road rate↓	0.008	0.015	
Collision rate↓	0.012	0.015	



Fig. 7: Evaluation of the ITRA-V-AIM on a Bing aerial image. The white trajectories represent the ground truth trajectories obtained from validation segments at the same location, and the orange trajectories display 5 sampled trajectories of the ITRA-V-AIM model. Our ITRA-V-AIM model predicts road context-aware trajectories on this aerial image which has different lighting and shading conditions compared to the training AIM.

proposed aerial image-based map (AIM) requires minimal annotations. By employing the AIM in a driving simulator through our image-texture-based differentiable rendering, we illustrate that the AIM provides rich road context information for multi-agent trajectory prediction which resulted in more realistic samples for both vehicles and pedestrians. Our results on pedestrian trajectory prediction indicates a substantial improvement when utilizing our AIM representation compared to the baseline representation. In addition, our image-texture-based differentiable rendering module can be easily integrated into any existing behavioral prediction models that consume bird's-eye view images as part of the agents' state representation. While our work has yielded promising results, there are still opportunities remaining for further improvement. These opportunities include an in-depth investigation of semantic or structure-based encoders for the AIM representation to improve on the off-road metrics. Finally, while we constructed AIMs from drone recordings, they could also be employed when collecting data from the vehicle by utilizing existing aerial imagery, although additional effort would be required to align such images to the extracted agent tracks.

VII. ACKNOWLEDGEMENT

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and the Mitacs Accelerate program in partnership with Inverted AI. Additional support was provided by UBC's Composites Research Network (CRN), and Data Science Institute (DSI). This research was enabled in part by technical support and computational resources provided by WestGrid (www.westgrid.ca), Compute Canada (www.computecanada.ca), and Advanced Research Computing at the University of British Columbia (arc.ubc.ca).

REFERENCES

- Y. Liu, J. W. Lavington, A. Scibior, and F. Wood, "Vehicle type specific waypoint generation," in *IEEE RSJ International Conference* on Intelligent Robots and Systems, 2022, pp. 12225–12230.
- [2] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "Trafficsim: Learning to simulate realistic multi-agent behaviors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10400–10409.
- [3] A. Ścibior, V. Lioutas, D. Reda, P. Bateni, and F. Wood, "Imagining the road ahead: Multi-agent trajectory prediction via differentiable simulation," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [4] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference* on Computer Vision, 2021, pp. 9710–9719.
- [5] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in *Conference on Robot Learning*. PMLR, 2021.
- [6] 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 mlbased planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.
- [7] M. Elhousni, Y. Lyu, Z. Zhang, and X. Huang, "Automatic building and labeling of hd maps with deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13 255–13 260.
- [8] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," 2022 International Conference on Robotics and Automation, pp. 4628–4634, 2021.
- [9] Y. Zhou, Y. Takeda, M. Tomizuka, and W. Zhan, "Automatic construction of lane-level hd maps for urban scenes," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2021, pp. 6649–6656.
- [10] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.
- [11] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kümmerle, H. Königshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," arXiv:1910.03088 [cs, eess], 2019.
- [12] X. Li, G. Li, Q. Huang, Z. Wang, and Z. Yu, "An adaptive background extraction method in traffic scenes," *Optik*, vol. 156, pp. 659–671, 2018.
- [13] R. Zhang, W. Gong, A. Yaworski, and M. Greenspan, "Nonparametric on-line background generation for surveillance video," in *Proceed*ings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE, 2012, pp. 1177–1180.
- [14] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions* on *Image processing*, vol. 20, no. 6, pp. 1709–1724, 2010.
- [15] O. Scheel, L. Bergamini, M. Wolczyk, B. Osiński, and P. Ondruska, "Urban driver: Learning to drive from real-world demonstrations using policy gradients," in *Conference on Robot Learning*, 2022, pp. 718– 728.
- [16] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in neural information processing systems*, vol. 28, 2015.

- [17] P. Polack, F. Altché, B. d'Andréa Novel, and A. de La Fortelle, "The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles?" in *IEEE Intelligent Vehicles Symposium (IV)*, 2017.
- [18] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings* of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011.
- [19] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 336–345.
- [20] Y. Tang and R. Salakhutdinov, "Multiple futures prediction," Advances in neural information processing systems, vol. 32, 2019.
- [21] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2821–2830.
- [22] X. Ma, J. K. Gupta, and M. J. Kochenderfer, "Normalizing flow policies for multi-agent systems," in *Decision and Game Theory for Security: 11th International Conference, GameSec 2020, College Park, MD, USA, October 28–30, 2020, Proceedings 11.* Springer, 2020, pp. 277–296.
- [23] Z. Zhong, D. Rempe, D. Xu, Y. Chen, S. Veer, T. Che, B. Ray, and M. Pavone, "Guided conditional diffusion for controllable traffic simulation," in 2023 IEEE International Conference on Robotics and Automation. IEEE, 2023, pp. 3560–3566.
- [24] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16.* Springer, 2020, pp. 541–556.
- [25] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 525–11 533.
- [26] K. Zhang, X. Feng, L. Wu, and Z. He, "Trajectory prediction for autonomous driving using spatial-temporal graph attention transformer," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [27] R. Korbmacher and A. Tordeux, "Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [28] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1349–1358.
- [29] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12126–12134.
- [30] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Socialstgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14424–14432.
- [31] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3d deep learning with pytorch3d," arXiv:2007.08501, 2020.
- [32] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664.
- [33] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2009, pp. 261–268.
- [34] "Bing Maps," [Accessed: April 12, 2023]. [Online]. Available: https://www.bing.com/maps
- [35] "Mobile Robot Kinematics," [Accessed: April 25, 2023]. [Online]. Available: https://stanfordasl.github.io/aa274a/pdfs/notes/lecture1.pdf
- [36] V. Lioutas, A. Scibior, and F. Wood, "TITRATED: Learned human driving behavior without infractions via amortized inference," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=M8D5iZsnrO