Semi-supervised Semantics-guided Adversarial Training for Robust Trajectory Prediction

Ruochen Jiao
Northwestern University
ruochen.jiao@u.northwestern.edu

Xiangguo Liu
Northwestern University
xg.liu@u.northwestern.edu

Takami Sato
University of California, Irvine
takamis@uci.edu

Qi Alfred Chen
University of California, Irvine
alfchen@uci.edu

Qi Zhu
Northwestern University
qzhu@northwestern.edu

Abstract

Predicting the trajectories of surrounding objects is a critical task for self-driving vehicles and many other autonomous systems. Recent works demonstrate that adversarial attacks on trajectory prediction, where small crafted perturbations are introduced to history trajectories, may significantly mislead the prediction of future trajectories and induce unsafe planning. However, few works have addressed enhancing the robustness of this important safety-critical task. In this paper, we present a ***novel adversarial training method for trajectory prediction****. Compared with typical adversarial training on image tasks, our work is challenged by more random input with rich context and a lack of class labels. To address these challenges, we propose a method based on a* ***semi-supervised*** *adversarial autoencoder, which models* ***disentangled semantic features*** *with domain knowledge and provides additional latent labels for the adversarial training. Extensive experiments with different types of attacks demonstrate that our* ***S****emi-supervised* ***S****emantics-guided* ***A****dversarial* ***T****raining (***SSAT***[1]) method can effectively mitigate the impact of adversarial attacks by up to 73% and outperform other popular defense methods. In addition, experiments show that our method can significantly improve the system's robust generalization to unseen patterns of attacks. We believe that such semantics-guided architecture and advancement on robust generalization is an important step for developing robust prediction models and enabling safe decision making.*

1. Introduction

Connected and autonomous vehicles (CAVs) have shown great promise to revolutionize the transportation systems, but also face significant concerns regarding their safety and security [15, 36, 21, 17, 50, 51]. Many of these concerns, which also apply to many other autonomous systems, arise from the increasing adoption of advanced deep neural network (DNN)-based machine learning techniques across system sensing, perception, prediction, planning, control, and general decision-making [52, 53].

In particular, the adversarial robustness of DNNs has drawn significant attention in recent years. State-of-the-art classifiers can be misled to make erroneous predictions by slight but carefully-optimized perturbations [39], and addressing this is critical to the safety of autonomous systems that employ these classifiers [11, 40]. Besides classification, adversarial attacks are also targeted at various other tasks such as object detection, image generation and natural language processing. To defend against adversarial examples, adversarial training is commonly used to enhance the intrinsic robustness of models and has been shown to be very effective among various defense strategies [24, 25, 34]. On the other hand, the works in [30, 43, 37] show that adversarial training may suffer poor robust generalization on unseen attacks and [33, 24, 38] demonstrate a trade-off between robustness and accuracy. In this paper, we address **trajectory prediction**, a safety-critical task that is common in self-driving vehicles and other autonomous systems, and propose a novel architecture to enhance its *adversarial robustness and robust generalization* by introducing semantic features and a semi-supervised adversarial autoencoder (AAE) [26] into the adversarial training process.

We focus on the trajectory prediction module in autonomous driving, where large DNNs have enabled breakthroughs in recent years. More specifically, an autonomous driving system typically consists of several modules such as perception, localization, prediction, planning (route, behavior, trajectory, and motion planning), and control. The perception module is to detect agents and obstacles based

[1] Our codes: https://github.com/jrcblue/SSAT-for-Motion-Prediction

on the sensing inputs (e.g., images and 3D point clouds), and the effectiveness and robustness of the perception module are well-studied in the research community. The prediction module is to predict *future trajectories* of surrounding vehicles based on the observed *history trajectories* of those vehicles from the perception module and the map context. The trajectory prediction plays a crucial role in understanding the environment and making safety-critical decisions in the following planning module [18, 20, 19]. Recent works [23, 16, 5, 45, 12] have applied various deep learning techniques such as graph neural networks and transformers to this task and achieved impressive performance in terms of reducing average errors. However, few works studied the robustness of vehicle trajectory prediction under adversarial attacks, which is in fact of vital importance because 1) autonomous driving is a safety-critical task by nature, 2) recent work [48, 10] demonstrates that the prediction module is vulnerable to adversarial attacks if surrounding vehicles drive along a seemingly natural but crafted trajectory, and 3) current prediction models are often overfitted on limited patterns in the datasets but suffer the long-tailed distribution of driving scenarios and behaviors. The threat model of the adversarial trajectory is illustrated in an example in Fig. 1, where a crafted input trajectory can intentionally mislead the target vehicle to the wrong prediction and induce dangerous planning decisions.

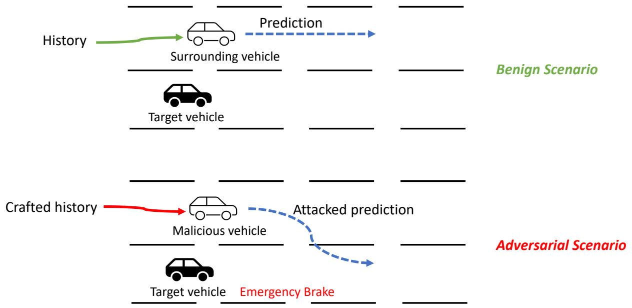


Figure 1: An illustrative example of the threat model: In the upper benign scenario, the target vehicle predicts the future trajectory of a surrounding vehicle based on its history trajectory. In the lower adversarial scenario, the surrounding vehicle is malicious and drives along a crafted (history) trajectory, which may mislead the target vehicle to have a wrong (attacked) prediction of the surrounding vehicle's future trajectory. In this particular case, the target vehicle wrongly predicts that the malicious vehicle will cut into its lane, and thus takes an unsafe emergency braking.

To defend against adversarial attacks on vehicle trajectory prediction, there are important challenges that are different from the cases of images and audios. First, vehicle trajectory prediction is a time-series regression problem with rich contexts while most existing adversarial attacks and corresponding defense methods are targeted at classification tasks. The attack patterns are more random and there are no well-defined class labels, which means that the robust model is difficult to train and generalize. Second, vehicle trajectory can convey semantic and behavior information. For instance, people can infer the behavior of a vehicle such as moving forward or changing lanes from its trajectory. Therefore, as shown in our experiments later, some popular defense methods such as TRADES [46] and data augmentation methods [6, 29, 35] either are inapplicable or have poor performance in trajectory prediction task. In this work, we first propose an adversarial training pipeline for the trajectory prediction task, and then further exploit the semantic features to design a semi-supervised AAE architecture that can be added after the feature extractor, to improve the adversarial robustness and its generalization. The methodology of the proposed architecture such as enhancing disentanglement and defining semantic labels may be further applied to adversarial training for general regression and generation problems. Our main contributions are summarized as follows:

- We propose a novel adversarial training method against adversarial attacks on trajectory predictions.
- We develop a semi-supervised architecture with domain knowledge and semantic features to enhance the adversarial robustness and its generalization among different types of attack patterns.
- Extensive experiments demonstrate that our proposed method (SSAT) effectively improves the adversarial robustness and outperforms popular defense baselines.
- We further explore to balance the robustness-accuracy trade-off in this task by leveraging the MixUp technique [47], and we also test data augmentation methods for trajectory prediction robustness.

2. Preliminaries

2.1. Adversarial Attacks on Trajectory Prediction

Recent research [48, 10] demonstrates that trajectory prediction in autonomous driving can be fooled by the adversarial behavior of a surrounding vehicle, where the adversarial behavior is optimized with Projected Gradient Descent (PGD) [24]. The threat model is illustrated in Fig. 1. In this work, we consider such attacks and assume the worst setting, i.e., the attacker has full knowledge of the target system and tries to maximize the attack impact. However, malicious agents cannot directly manipulate anything inside the target vehicle and they can only slightly change their own trajectories to mislead the target vehicle indirectly.

Different from image classification, trajectory prediction has no class labels, but it has directional information in the

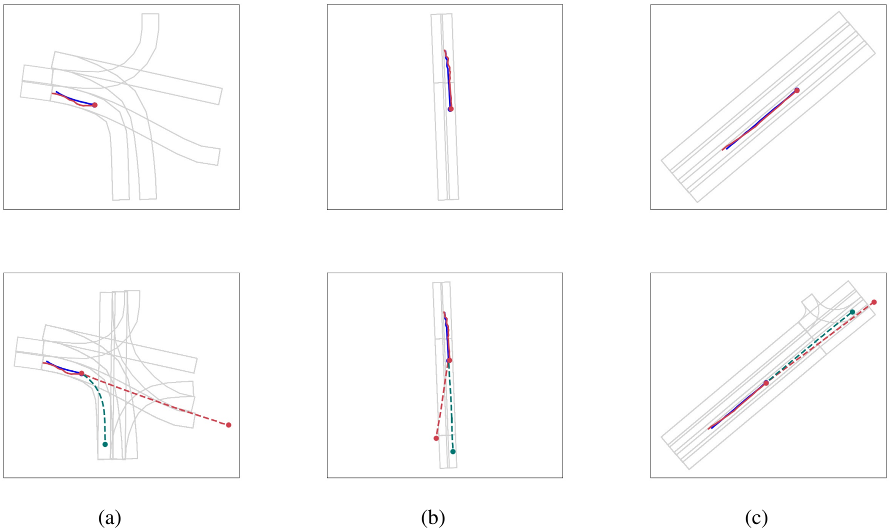


Figure 2: Adversarial trajectories generated by different types of attacks. The figures in the top row are adversarial (red line) and benign (blue line) input history trajectories for prediction and they look very close to human eyes. The figures in the bottom row show the corresponding attacked future trajectory prediction (red dashed line) and the ground truth trajectory in the benign case (green dashed line). The differences between the two are clearly visible, showing the effectiveness of the attacks. Figure (a) is the ADE attack, which will randomly lead to maximum average deviation. Figure (b) is the lateral attack, which mainly leads the vehicle to deviate to the left or right. Figure (c) is the longitudinal attack, which will mainly lead to longitudinal deviation.

context, such as moving forward, turning or changing lanes to the right. Therefore, the attacker can conduct targeted attacks besides random ones. In Fig. 2, three types of adversarial attacks are presented and we observe that they can cause significant and directional errors. [48] proposes directional error metrics for the optimization of targeted attacks, as shown in Eq. (1):

$$D(\alpha, R) = (p_\alpha - s_\alpha)^T \cdot R\,(s_{\alpha+1}, s_\alpha)\,, \qquad (1)$$

where α denotes the time frame ID. p and s are two-dimensional vectors representing predicted and ground-truth vehicle locations, respectively. R is a function generating the unit vector to a specific direction (lateral or longitudinal). The longitudinal direction is approximated as the vector defined by the adjacent two waypoints of ground truth $s^n_{\alpha+1} - s^n_\alpha$.

In addition to the directional attack, a random attack can be designed to maximize the Average Displacement Error (ADE), which is the average of the root mean squared error between the predicted waypoints and the ground-truth trajectory waypoints.

Hard constraints are also applied to the maximum deviation of waypoints during the optimization so as to make the adversarial history trajectory physically feasible and not perform unrealistic behaviors.

Table 1: Trajectory prediction errors (in meters) in benign or adversarial scenarios under different types of attacks.

Model	ADE	Lateral	Longitudinal
	benign/attack	benign/attack	benign/attack
LaneGCN [16]	1.32/5.17	−0.01/1.58	−0.25/3.33
Trajectron++[31]	2.69/6.81	0.107/2.25	-0.526 / 3.79

Table 1 quantitatively shows that the predicted trajectories can be seriously deviated by any of the three attack types. Generally, in the US, 0.3m lateral deviation is enough to invade adjacent lanes on local roads [32, 8]. In addition, the final displacement error (the distance between the last points of predicted and ground-truth trajectories) is around two to three times larger than the average error shown here. Therefore, the attacker can apply either random or directional attacks that will greatly challenge the defense methods, especially on the generalization performance.

[48, 2] propose to utilize approaches such as smoothing and adversarial training on the conditional variational encoder (CVAE) architectures to mitigate the impact of such adversarial attacks. However, they do not explicitly address driving semantics and robustness generalization, which are critical for many real road scenarios.

2.2. Adversarial Training Methods and Robust Generalization

Adversarial training [24, 25, 34] is shown to be one of the most effective approaches to improve the robustness of DNN models. In practice, the PGD attack is commonly deployed for evaluation because of its strong attack ability in white-box settings and the work in [24] formulates the adversarial training as a min-max problem. [6, 29, 35] show that synthesized data can significantly boost robustness for the image classification task. Thus, we select similar methods as baselines for comparison in our experiments.

Recent research shows that a specific type of adversarial attack is not sufficient to represent the diverse space of adversarial examples and many adversarially trained models are only robust to specific attacks. This does limit the application of adversarial training in practice, especially for trajectory regression tasks due to its nature of long-tailed distribution. The works in [33, 49, 37] try to explain and improve the robust generalization in various perspectives such as sample complexity and latent feature representation. In this work, we demonstrate that robust generalization for trajectory prediction can be enhanced by explicitly introducing disentangled and semantic features in the latent space.

2.3. Adversarial Autoencoder Architecture

The adversarial autoencoder (AAE) is a variant of the variational autoencoders (VAE) [14, 13], which provides a principled method for jointly learning deep latent-variable models and corresponding inference models using stochastic gradient descent. AAE imposes various distributions on the latent vector by utilizing adversarial learning instead of KL divergence. Due to the flexibility of adversarial learning, AAE is superior to VAE in terms of imposing complicated distributions over latent space. For adversarial robustness, the works in [27, 41] demonstrate that the disentangled latent representations produce VAEs that are more robust to adversarial attacks. In our work, we design an AAE-based architecture that can be added after the feature extractor of prediction modules. We utilize this architecture to model diverse semantic features and enhance disentanglement in the latent space.

3. Proposed Adversarial Training Method for Trajectory Prediction

3.1. Domain Knowledge-guided Semi-supervised Architecture

3.1.1 Overall Design

Unlike adversarial training for image classification tasks, there are no class labels to guide the training for trajectory prediction, and the attack patterns in the trajectory regression are more random. We thus introduce domain knowledge to facilitate the modeling of the semantic information in both benign and adversarial cases, based on an AAE architecture. The model learns the directional semantic latent vector in a semi-supervised way because the ground truths are only available for limited scenarios but their distributions can be derived from domain knowledge and statistics. Therefore, our latent space modeling contains two levels – unsupervised distribution modeling and semi-supervised learning when labels are available.

The architecture of our proposed model is shown in Fig. 3. The feature extractor [16] utilizes a one-dimensional dilated convolutional neural network to obtain the embedding of the time-series trajectory and uses a graph neural network to model the lane context and interaction between objects. The encoder maps high-dimensional features to the semantics-guided latent space with distribution regularization and semi-supervised training. In particular, the latent space is divided into three parts: longitudinal features z_{lon}, following one-dimensional log-normal distribution, lateral features z_{lat}, following three-dimensional categorical distribution and remaining features, following Gaussian distribution [12]. Finally, the decoder maps semantic vectors along with other disentangled latent vectors to the future trajectories. Note that we develop the AAE instead of traditional VAE architecture to model these different and complex distributions.

In the attacked scenarios, the impact of attacks will be decomposed to different latent vectors and the attack patterns will be explicitly modeled by the semantic features. Let us take the lateral directional vector as an example. If the attack is not targeted at the lateral dimension, the encoder will decompose the attack effect into other vectors and the mapping for lateral direction will remain stable. Otherwise, if the attack causes errors in the lateral vector, the feature extractor and encoder will be adversarial trained on the label of the direction and the corresponding mapping from the latent distribution to the final trajectories will also be updated.

Compared to existing trajectory prediction works that only apply adversarial training on the final trajectory waypoints, our proposed method is designed to capture semantic features in the latent space, and it can benefit the adversarial robustness and its generalization in various aspects:

- The architecture maps the high-dimensional features into disentangled latent space that can decompose the attacks into different orthogonal patterns. Such disentanglement will boost the adversarial robustness [41].

- The semantic vectors provide more context information and interpretable labels for the adversarial training and the semi-supervised learning makes the adversarial training process more efficient and easier to generalize to the unseen scenarios.

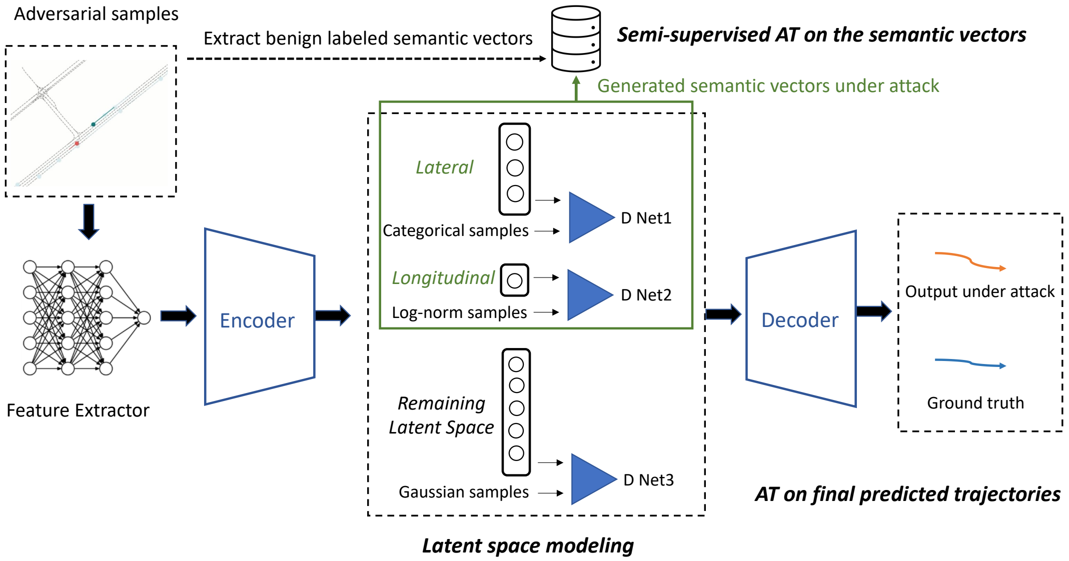


Figure 3: The overall adversarial training pipeline for the proposed semi-supervised semantics-guided architecture. We leverage the feature extractor in [16]. In the latent space, the D nets represent the discriminator that models the latent distributions by distinguishing the samples of real distribution from the predicted ones. When the semantic labels are available in the scenarios, they are used to adversarially train the semantic vectors (parts in green). Finally, we conduct adversarial training on attacked trajectories.

- The encoder-decoder architecture can filter high-frequency noise by dimension reduction.

The detailed design and optimization of the architecture are described in the following.

3.1.2 Semi-supervised Semantic Features Modeling

To model the high-level semantic information in the trajectory prediction task, it is natural to decompose the trajectories into two dimensions – longitudinal and lateral directions. We want to utilize domain knowledge to guide the modeling by providing representative metrics and prior distributions. In addition to the semantic features, the architecture also maps high-dimensional information into latent vectors in Gaussian distributions to represent other low-level and more random features. We will mainly explain the semantic feature modeling below.

In the longitudinal direction, speed and acceleration are often used to model the vehicle's dynamics but their values are always changing and do not contain enough semantic information. In our model, we apply the *time headway* to effectively extract the longitudinal feature, which measures the time difference between two successive vehicles when they cross a given point. The time headway represents a relatively stable behavior pattern for a certain agent and takes interaction with other vehicles into consideration. Recent works [44, 22] also use time headway as a measurement of aggressiveness in specific scenarios such as lane changing or merging. In our proposed architecture, the model will represent the time headway as a one-dimensional vector in the latent space. In both benign and adversarial cases, the encoder will be trained by the regularization loss to force the longitudinal feature to a certain distribution that the time headway follows in statistics. Prior works [7] in the transportation domain show that the time headway in urban scenarios can be best described with the Log-normal distribution. We verify this and estimate the parameters in the Argoverse 1 motion forecasting dataset [4]. The longitudinal feature follows the distribution shown in Eq. (2):

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}, \qquad (2)$$

where μ is the location parameter and σ is the scale parameter. We can explicitly obtain the true time headway values for semi-supervised training when there is observable interaction between the attacked target and the front vehicle. We consider the semi-supervised longitudinal feature encoding as a regression problem and optimize it by minimizing the mean square error. For the lateral directional features, we represent them by three simple but effective classes: moving forward, turning/changing lanes to the left, and turning/changing lanes to the right. These three intentions are discrete by nature and we model them with the categorical distribution. In the adversarial training process, only vehicles with clear intentions in a long enough time frame will be labeled and we utilize the cross-entropy to optimize this classification task.

For all the semantic and Gaussian latent variables, they are regularized to the target distribution by the adversarial generation loss in Eq. (3). The discriminators are trained to maximize the log probability of real latent samples s and the log of the inverse probability for fake latent samples, as in Eq. (4):

$$Loss_G(x) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1 - D_i\left(G\left(x\right)\right)\right), \tag{3}$$

$$Loss_{D_i}(x, s) = \log D_i\left(s_i\right) + \log\left(1 - D_i\left(G\left(x\right)\right)\right), \tag{4}$$

where x is the high-dimensional features and m is the number of different kinds of latent vectors. G and D are encoders and distribution discriminators, respectively.

3.2. Adversarial Training Process

3.2.1 Adversarial Training Algorithm

For each sample, we utilize the PGD attack to generate the adversarial trajectory only for the target vehicle and keep other surrounding vehicles' original trajectories. This constrains the adversarial attack's impact on the whole scenario. If the error between the prediction under attack and the ground truth is greater than a threshold, we consider the attack a successful one and conduct adversarial training on this sample. Since the perturbation is very small, we consider the real future trajectory as the ground truth y_i for the adversarial training and optimize the whole pipeline with L1-smooth loss in Eq. (5). This is a general but somewhat naive adversarial training process.

$$Loss_{traj}\left(y_i, \hat{y}_i\right) = \begin{cases} 0.5\left(y_i - \hat{y}_i\right)^2 & \text{if } \|y_i - \hat{y}_i\| < 1, \\ \|y_i - \hat{y}_i\| - 0.5 & \text{otherwise}. \end{cases} \tag{5}$$

Thus, to further facilitate the adversarial training, we exploit semantic features and their corresponding labels in our proposed architecture. The encoder is optimized to minimize the mean square error of longitudinal features and the cross entropy of lateral features between the ground truths and predictions in the latent space. The semi-supervised loss function is shown as follows:

$$Loss_{Semi}(z, g) = -\sum_{i=1}^{3} g_{lat} \log z_{lat} + (g_{lon} - z_{lon})^2, \tag{6}$$

where z represents the predicted semantic vectors when under attack and g represents the ground truth in the benign scenarios.

Moreover, we further adapt the adversarial training process with lateral semantic vectors because the lateral directional prediction can be regarded as a classification problem with clear behavior meaning. When the adversarial example leads to the wrong classification of lateral behavior, we will set higher weights of the semi-supervised loss for the adversarial training. In this way, our model will first guarantee the correctness of high-level semantic prediction and then tune the regression error, which could help avoid significant adversarial deviation and enhance the generalization performance.

Algorithm 1 Our Adversarial Training Pipeline

```
Initialize: feature extractor $F$, AAE encoder $G$, decoder $R$, discriminator $D_i$, target distribution $p_i$, $i$ = 1,2,3, adversarial example generator $Adv$.
Input: past trajectories $t$, future trajectories ground truth $y$ , map context $c$, prediction model $m$.
for each sample do
    Generate adversarial trajectory for target vehicle $t_{adv} = Adv(t, y, m, c)$.
    Input $t_{adv}$ to the predictor. $\hat{y}', \hat{z}' = m(t_{adv}, c)$, where $z$ =$\{z_{lat}, z_{lon}, z_{gaussian}\}$.
    if $err(\hat{y}', y) > Threshold$ then
        if $argmax(z_{lat}) \neq argmax(\hat{z_{lat}}')$ then
            Update model $m$ with higher weights of semi-supervised learning loss $Loss_{semi}$ in Eq. (6).
            Update model $m$ with $Loss_D$ in Eq. (4), $Loss_G$ in Eq. (3), and $Loss_{traj}$ in Eq. (5).
        else
            Update model $m$ with $Loss_D$ in Eq. (4), $Loss_G$ in Eq. (3), $Loss_{semi}$ in Eq. (6), and $Loss_{traj}$ in Eq. (5).
        end if
    end if
end for
```

3.2.2 Balance Accuracy and Robustness

In our preliminary experiments, we notice a trade-off between standard accuracy and adversarial robustness. A similar phenomenon has been observed in classification tasks [46, 28]. Methods such as TRADES [46], robust self-training [3] and MixUp [47, 1] have been proposed to balance such trade-off. However, there are few methods that can be applied to trajectory prediction because such time-series regression problems have no class labels and are more sensitive to errors introduced by augmented data. In this work, we utilize the MixUp [47] technique to mix the adversarial scenarios and benign scenarios in the adversarial training process. The experiments demonstrate that a balance between adversarial robustness and standard accuracy can be achieved in trajectory prediction, as shown later.

Table 2: Comparison of different defense methods when under various attacks and in the benign case. For the adversarial training methods, we calculate the mean error of models adversarially trained on different attacks.

Methods	ADE Attack			Lateral Attack			Longitudinal Attack			Benign		
	ADE/m			Lateral Error/m			Longitudinal Error/m			ADE/m		
Dataset	Argo1	Argo2	Apollo	Argo1	Argo2	Apollo	Argo1	Argo2	Apollo	Argo1	Argo2	Apollo
Original Model	5.17	4.52	3.93	1.66	1.27	1.54	3.78	3.17	3.42	**1.43**	**0.79**	**1.78**
Train-time Smo. [48]	4.67	4.32	3.86	1.45	1.27	1.49	3.23	3.00	3.41	1.50	0.83	2.11
Test-time Smo. [48]	4.32	3.51	3.49	0.75	0.89	1.09	3.54	2.42	3.05	1.68	1.23	2.07
Heuristic Aug. [29]	4.62	3.52	3.84	1.09	1.21	1.34	3.24	4.24	3.21	1.54	0.97	2.01
Data-driven Aug. [29]	4.53	4.43	3.53	0.74	0.78	0.54	2.73	2.61	3.02	2.50	1.99	2.24
Standard AT	3.79	3.67	3.68	0.67	0.64	1.15	2.44	1.89	3.05	1.67	1.06	1.87
SSAT (ours)	**3.51**	**2.67**	**2.90**	**0.61**	**0.53**	**0.41**	**1.76**	**1.44**	**1.26**	1.75	1.20	1.87
Mixup-SSAT (ours)	3.73	3.33	3.17	0.72	0.51	0.53	2.13	1.63	1.35	1.64	1.07	1.86

4. Experimental Results

4.1. Experiment Setup

4.1.1 Dataset

We train and evaluate different defense methods using three popular benchmarks – Argoverse 1 [4], Argoverse 2 [42], and ApolloScape datasets [9]. The datasets contain more than 250k real driving scenarios in different cities, such as Miami and Pittsburgh. For Argoverse 1 and Argoverse 2, each scenario consists of a road graph and multiple agents' trajectories sampled at a frequency of 10Hz. We choose 20 waypoints as the history trajectory and the models will predict 30 waypoints in the future. Scenarios in Apolloscape are simpler. They have no maps but 6 waypoints for both history and future trajectories.

4.1.2 Attack Settings

In the experiments, we study three different types of attacks [48] to the vehicle trajectory prediction algorithms – lateral directional attack (shift to the right), longitudinal directional attack (shift forward), and ADE attack (deviate randomly). More details of them are in Section 2.1. The trajectory prediction models are adversarially trained on the three types of attacks, respectively. We constrain the maximum deviation between the attacked and the benign input trajectories to be 1 meter.

4.1.3 Training Settings

Since our architecture is an encoder-decoder module that can be combined with different feature extractors, we first fine-tune the model on the benign data. In the experiments, we use the feature extractor from LaneGCN [16], an attention-based graph neural network. We notice that the AAE architecture introduces a slight accuracy drop on the benign data, mainly due to the dimension reduction. For adversarial training, we train prediction models on adversarial samples generated from scenarios in datasets.

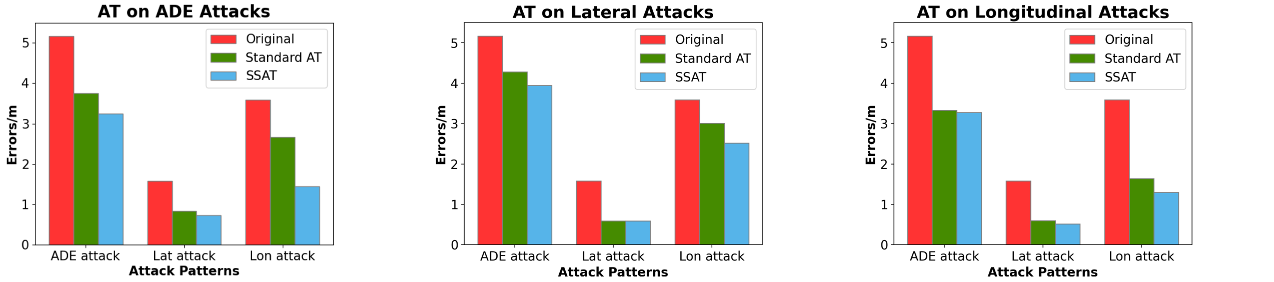
4.2. Experimental Results and Analysis

In this section, we conduct experiments with various defense methods under different patterns of attacks, including our semi-supervised semantics-guided (SSAT) method and a Mixup-SSAT method that combines SSAT with the MixUp technique for balancing standard accuracy and adversarial robustness. In the following, we first compare the average robustness improvement among various defense methods, which demonstrates the advantage of SSAT in improving robustness under various types of attacks and the effectiveness of Mixup-SSAT in balancing robustness and accuracy. Then, we show that SSAT can significantly enhance the robust generalization to unseen types of attacks. In addition, we evaluate an unsupervised version of SSAT to explicitly show how the semi-supervised semantic-guided latent space modeling can boost the adversarial robustness, which also serves as an ablation study.

4.2.1 Effectiveness of SSAT Methods

We compare *our SSAT and Mixup-SSAT methods with the original model and five different defense methods*, including train-time smoothing [48], test-time smoothing [48], heuristic data augmentation [29], data-driven augmentation [29], and the standard adversarial training (Standard AT). All the models share the same feature extractor LaneGCN in this setting. Note that we compare to the two data augmentation methods as they are effective for image classification tasks [6, 29, 35]. For the data-driven augmentation, we design an additional decoder to augment input trajectories and it can generate more inputs by adding Gaussian noises to the latent vectors of the real inputs. For the heuristic augmentation, we simply add random perturbation to the benign inputs, with the same constraints of maximum deviation.

Table 2 shows the prediction errors of different methods when under various types of attacks and in the benign case (note that when under lateral and longitudinal attacks, we measure lateral and longitudinal errors). We can see that **our SSAT method significantly outperforms all other de-**



(a) (b) (c)

Figure 4: Adversarial robustness comparisons of the original model, the standard adversarial training on final trajectories (Standard AT), and the SSAT method proposed by us. Figures (a), (b), (c) show results from adversarial training (AT) that targets ADE attacks, lateral attacks, and longitudinal attacks, respectively.

fense methods in improving the robustness of trajectory prediction. Compared with the original model, SSAT can reduce the prediction error by $32\% - 73\%$ when under different types of attacks.

Moreover, while our SSAT method improves robustness, we also observe a drop in standard accuracy in the benign case. **Mixup-SSAT enables effective trade-off between these two objectives** (i.e., better performance than SSAT in benign cases at the expense of worse performance under attacks), by setting the mixup ratio between adversarial and benign examples to different values (the results in Table 2 are based on a mixup ratio of 2).

We also notice that both data-driven and heuristic data augmentation methods offer very limited improvement over the original model. This is likely due to the challenge of regression tasks with rich context, which makes augmentation methods perform poorer than they do for image classification tasks.

Table 3: Comparison of different methods when adversarial training is conducted for **ADE** attack, and tested on ADE, lateral, and longitudinal attacks.

Methods	Prediction Error/m, w/o AT → with AT		
	ADE Attack	Lat Attack	Lon Attack
Standard AT	5.17 → 3.76	1.58 → 0.83	3.59 → 2.66
SSAT	5.16 → **3.24**	1.66 → **0.73**	3.78 → **1.45**
Unsup-SSAT	5.16 → 3.48	1.66 → 0.74	3.78 → 1.49

Table 4: Comparison of different methods when adversarial training is conducted for **lateral** attack, and tested on ADE, lateral, and longitudinal attacks.

Methods	Prediction Error/m, w/o AT → with AT		
	ADE Attack	Lat Attack	Lon Attack
Standard AT	5.17 → 4.28	1.58 → **0.59**	3.59 → 3.01
SSAT	5.16 → **4.02**	1.66 → **0.59**	3.78 → **2.52**
Unsup-SSAT	5.16 → 4.11	1.66 → 0.69	3.78 → 2.54

4.2.2 Effectiveness of SSAT in Robust Generalization on Different Types of Attacks

We observe that there is an adversarial robust generalization gap when the training and test are under different types of attacks. The comparisons in Tables 3, 4, 5 show that **our SSAT method is better generalized to unseen types of attacks**, compared to the standard adversarial training. In every training scenario, our method is more robust to the various unseen patterns of attacks. For instance, Table 3 shows that when applying SSAT to train under the random ADE attack, its results outperform other models on all seen (i.e., ADE) and unseen (i.e., lateral and longitudinal) types of attacks, which demonstrates that our SSAT method can effectively decompose and learn semantic features from random ADE attacks. Tables 4 and 5 show similar trends, where our SSAT methods are better at defending against unseen attacks and mitigating overfitting on specific patterns of attacks. Fig. 4 further visualizes the results from these three tables for the original model, the standard adversarial training (Standard AT), and SSAT.

Table 5: Comparison of different methods when adversarial training is conducted for **longitudinal** attack, and tested on ADE, lateral, and longitudinal attacks.

Methods	Prediction Error/m, w/o AT → with AT		
	ADE Attack	Lat Attack	Lon Attack
Standard AT	5.17 → 3.33	1.58 → 0.60	3.59 → 1.64
SSAT	5.16 → **3.28**	1.66 → **0.52**	3.78 → **1.30**
Unsup-SSAT	5.16 → 3.40	1.66 → 0.61	3.78 → 1.49

4.2.3 Impact from Latent Space Modeling

We also conduct adversarial training that only has regularization on the latent distributions but without supervision on latent vectors. We name it Unsup-SSAT. The comparisons between the standard adversarial training (Standard AT) and Unsup-SSAT in Tables 3, 4, and 5 demonstrate that even without labels, the partial disentanglement and distribution modeling in Unsup-SSAT will benefit the adversarial training on trajectory prediction and outperform the baseline Standard AT in most cases. However, when compared with SSAT, we find that the adversarial robustness will be further improved with the extra labels in the semi-supervised phase (in practice, we often have access to those labels).

5. Conclusion

In this work, we propose an adversarial training method for trajectory prediction. To tackle the challenge of random inputs with rich context, diverse types of attacks, and lack of class labels, we develop a novel AAE architecture that exploits the disentanglement and semantic features for enhancing model robustness and its generalization. Our proposed SSAT method significantly outperforms a number of baselines from the literature, reducing the prediction errors under attacks by up to $32\% - 73\%$ when compared with the original prediction models. Our method is also shown to be effective for defending against unseen attacks.

Acknowledgment

We gratefully acknowledge the support from NSF grants 1834701, 1724341, 2038853, 1929771, 2145493, 1932464, and ONR grant N00014-19-1-2496.

References

[1] Guillaume P Archambault, Yongyi Mao, Hongyu Guo, and Richong Zhang. Mixup as directional adversarial training. *arXiv preprint arXiv:1906.06875*, 2019.

[2] Yulong Cao, Danfei Xu, Xinshuo Weng, Zhuoqing Mao, Anima Anandkumar, Chaowei Xiao, and Marco Pavone. Robust trajectory prediction against adversarial attacks. In *6th Annual Conference on Robot Learning*.

[3] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.

[5] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020.

[6] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

[7] Duy-Hung Ha, Maurice Aron, and Simon Cohen. Time headway variable and probabilistic modeling. *Transportation Research Part C: Emerging Technologies*, 25:181–201, 2012.

[8] Michael W Hancock and Bud Wright. A policy on geometric design of highways and streets. *American Association of State Highway and Transportation Officials: Washington, DC, USA*, 2013.

[9] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018.

[10] Ruochen Jiao, Juyang Bai, Xiangguo Liu, Takami Sato, Xiaowei Yuan, Qi Alfred Chen, and Qi Zhu. Learning representation for anomaly detection of vehicle trajectories. 2023.

[11] Ruochen Jiao, Hengyi Liang, Takami Sato, Junjie Shen, Qi Alfred Chen, and Qi Zhu. End-to-end uncertainty-based mitigation of adversarial attacks to automated lane centering. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 266–273. IEEE, 2021.

[12] Ruochen Jiao, Xiangguo Liu, Bowen Zheng, Dave Liang, and Qi Zhu. Tae: A semi-supervised controllable behavior-aware trajectory generator and predictor. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12534–12541. IEEE, 2022.

[13] Diederik Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Dasom Lee and David J Hess. Regulations for on-road testing of connected and automated vehicles: Assessing the potential for global safety harmonization.

Transportation Research Part A: Policy and Practice, 136:85–98, 2020.

[16] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556. Springer, 2020.

[17] C. Lin, Bowen Zheng, Qi Zhu, and Alberto Sangiovanni-Vincentelli. Security-Aware Design Methodology and Optimization for Automotive Systems. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 21(1):18:1–18:26, December 2015.

[18] Xiangguo Liu, Chao Huang, Yixuan Wang, Bowen Zheng, and Qi Zhu. Physics-aware safety-assured design of hierarchical neural network based planner. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*, pages 137–146. IEEE, 2022.

[19] Xiangguo Liu, Ruochen Jiao, Yixuan Wang, Yimin Han, Bowen Zheng, and Qi Zhu. Safety-assured speculative planning with adaptive prediction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[20] Xiangguo Liu, Ruochen Jiao, Bowen Zheng, Dave Liang, and Qi Zhu. Safety-driven interactive planning for neural network-based lane changing. In *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, pages 39–45, 2023.

[21] Xiangguo Liu, Yunpeng Luo, Anthony Goeckner, Trishna Chakraborty, Ruochen Jiao, Ningfei Wang, Yixuan Wang, Takami Sato, Qi Alfred Chen, and Qi Zhu. Invited: Waving the double-edged sword: Building resilient cavs with edge and cloud computing. In *Proceedings of the 60th Annual Design Automation Conference*, 2023.

[22] Xiangguo Liu, Neda Masoud, and Qi Zhu. Impact of sharing driving attitude information: A quantitative study on lane changing. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1998–2005. IEEE, 2020.

[23] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020.

[26] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[27] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019.

[28] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

[29] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

[30] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[31] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700. Springer, 2020.

[32] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3309–3326, 2021.

[33] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[34] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

[35] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021.

[36] Amolika Sinha, Sai Chand, Kasun P Wijayaratna, Navreet Virdi, and Vinayak Dixit. Comprehensive safety assessment in mixed fleets with connected and automated vehicles: A crash severity and rate evaluation of conventional vehicles. *Accident Analysis & Prevention*, 142:105567, 2020.

[37] Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E Hopcroft. Robust local features for improving the generalization of adversarial training. *arXiv preprint arXiv:1909.10147*, 2019.

[38] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[40] Zhilu Wang, Chao Huang, and Qi Zhu. Efficient global robustness certification of neural networks via interleaving twin-network encoding. In *DATE'22: Proceedings of the Conference on Design, Automation and Test in Europe*, 2022.

[41] Matthew Willetts, Alexander Camuto, Tom Rainforth, Stephen Roberts, and Chris Holmes. Improving vaes' robustness to adversarial attack. *arXiv preprint arXiv:1906.00230*, 2019.

[42] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[43] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[44] Hongtao Yu, H Eric Tseng, and Reza Langari. A human-like game theory-based controller for automatic lane changing. *Transportation Research Part C: Emerging Technologies*, 88:140–158, 2018.

[45] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.

[46] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[47] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[48] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2022.

[49] Shufei Zhang, Zhuang Qian, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Xinping Yi. Towards better robust generalization with shift consistency regularization. In *International Conference on Machine Learning*, pages 12524–12534. PMLR, 2021.

[50] Bowen Zheng, Hengyi Liang, Qi Zhu, Huafeng Yu, and Chung-Wei Lin. Next generation automotive architecture modeling and exploration for autonomous driving. In *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 53–58, 2016.

[51] Bowen Zheng, Chung-Wei Lin, et al. Design and Analysis of Delay-Tolerant Intelligent Intersection Management. *ACM T-CPS*, 4(1):3:1–3:27, November 2019.

[52] Qi Zhu, Chao Huang, Ruochen Jiao, Shuyue Lan, Hengyi Liang, Xiangguo Liu, Yixuan Wang, Zhilu Wang, and Shichao Xu. Safety-assured design and adaptation of learning-enabled autonomous systems. In *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, 2021.

[53] Qi Zhu, Wenchao Li, Hyoseung Kim, Yecheng Xiang, Kacper Wardega, Zhilu Wang, Yixuan Wang, Hengyi Liang, Chao Huang, Jiameng Fan, and Hyunjong Choi. Know the unknowns: Addressing disturbances and uncertainties in autonomous systems. In *Proceedings of the 39th International Conference on Computer-Aided Design*, ICCAD '20, New York, NY, USA, 2020. Association for Computing Machinery.