Supplementary Material: 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

In the following Supplementary Materials, we provide additional experimental results on more attack patterns in Section "Experiments with More Unseen Types of Adversarial Attacks", and detailed results of SSAT and Mixup SSAT on benign data in Section "Detailed Results on Benign Data". In Section "More Data-efficient Training for SSAT", we demonstrate that our SSAT method can achieve better performance with even fewer adversarial examples than the standard adversarial training. Section "Detailed Design of the Model" and Section "Details on Experiment Settings" provide details on model design and training settings, respectively. Finally, we discuss the limitations and future work in Section "Limitations and Future Work".

1. Experiments with More Unseen Types of Adversarial Attacks

In the experiment section of the main paper, subsection 4.2.2 "Effectiveness of SSAT in Robust Generalization on Different Types of Attacks", we study the robustness of our SSAT approach and the standard adversarial training method (Standard-AT) against various types of attacks including those targeting ADE (Average Displacement Error), lateral deviation and longitudinal deviation. This is motivated by the observation from recent works [3, 1] that there is a gap in robust generalization, i.e., the adversarially trained model may be robust to a specific type of attack but can be circumvented easily by other types of attacks (that are unseen during training), e.g., other l_p norms, different attack targets, or different constraints such as ϵ or maximum deviation in trajectory prediction.

In this supplementary material, we further evaluate the robustness of our SSAT approach and the standard adversarial training method (Standard-AT) against two additional types of attacks, **L1 and FDE**, with different l_p norm, larger constraints, and one more target function. For these two attack types, we relax the maximum deviation constraint from 1m to 1.5m. The experiments shown below on these two additional attack types L1 and FDE further demonstrate that **our proposed SSAT method has better robust generalization than the standard adversarial training method (Standard-AT)**. The experiments are conducted in Argoverse 1 dataset[2].

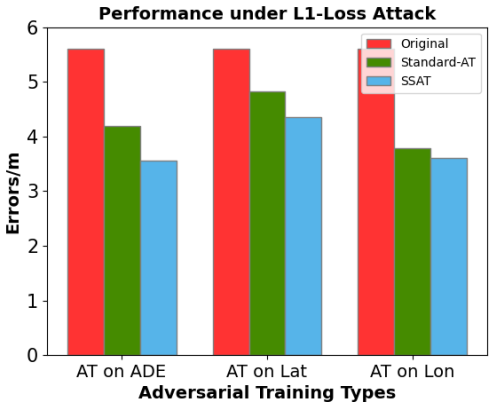


Figure 1: **L1 attacks:** We apply adversarial examples optimized for l_1 loss to attack the models that are adversarially trained on ADE, lateral and longitudinal attacks (denoted as "AT on ADE", "AT on Lat", "AT on Lon", respectively). The result shows that the models trained with our SSAT method (on any one of ADE, lateral and longitudinal attacks) are more robust to the unseen L1 attacks than the original model (Original) and the standard adversarial training method (Standard-AT).

1.1. L1 Attacks

We optimize the adversarial examples to maximize the l_1 loss, as shown below in Equation (1), between the pre-

dicted trajectories and the ground truth. Then, we attack the models that are adversarially trained on ADE, lateral and longitudinal attacks, respectively. The performance is measured with the average displacement errors.

$$\sum_{i=1}^{n} |y_{true} - y_{predicted}| + \sum_{i=1}^{n} |x_{true} - x_{predicted}| \quad (1)$$

The results in Fig. 1 demonstrate that the model adversarially trained with our approach SSAT is more robust to L1 attacks than the standard adversarial training method (Standard-AT).

1.2. FDE Attacks

Final Displacement Error (FDE) measures the root mean squared error (RMSE) of the last waypoints between the predicted trajectories and the ground truth. We use it as a new target for optimizing the adversarial examples and then attack the models that are adversarially trained on ADE, lateral, and longitudinal attacks, respectively. The results are shown in Fig. 2. SSAT again shows better performance (i.e., more robust) on this new type of attacks than Standard-AT.

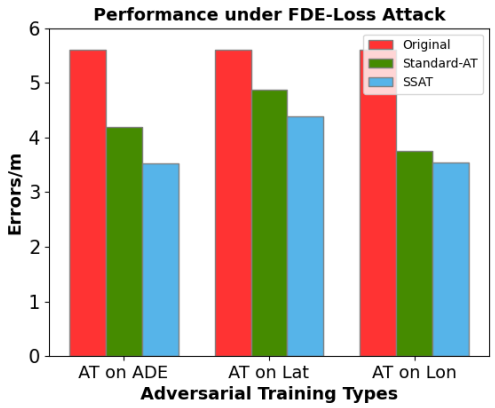


Figure 2: **FDE attacks:** We apply adversarial examples optimized for FDE to attack the models that are adversarially trained on ADE, lateral and longitudinal attacks (denoted as "AT on ADE", "AT on Lat", "AT on Lon", respectively). The result shows that the models trained with our SSAT method (on any one of ADE, lateral and longitudinal attacks) are more robust to the unseen FDE attacks than the original model (Original) and the standard adversarial training method (Standard-AT).

2. Detailed Results on Benign Data

In Table 1, we provide more detailed results of the proposed adversarial training methods' performance on benign examples in the Argoverse 1 dataset, measured by ADE (Average Displacement Error), FDE (Final Displacement Error), and MR (missing rate). Note that MR is the ratio of predictions whose final position is more than 2 meters away from the ground truth. We can find that our SSAT method trained under various attacks could lead to some reduction in standard accuracy for benign data. By applying the MixUp technique, the accuracy drop on benign data can be mitigated, especially for ADE attacks.

Table 1: We present the performance of our proposed methods on benign samples. The models are adversarially trained on different patterns of attacks and evaluated with three metrics.

Methods	ADE	FDE	MR
Original Model	1.43	3.08	0.53
SSAT on ADE attack	1.85	3.89	0.64
Mixup-SSAT on ADE attack	1.62	3.50	0.58
SSAT on Lat attack	1.65	3.61	0.62
MixUp-SSAT on Lat attack	1.64	3.61	0.62
SSAT on Lon attack	1.65	3.61	0.62
MixUp-SSAT on Lon attack	1.67	3.58	0.60

3. More Data-efficient Training for SSAT

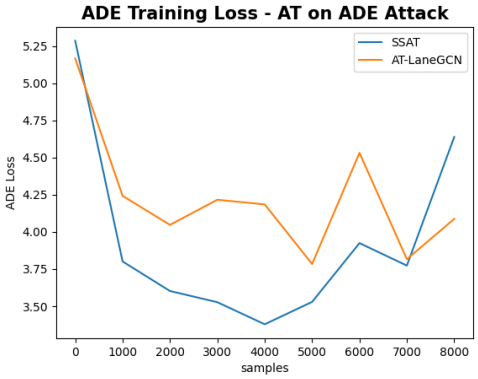
Fig. 3 demonstrates that our SSAT method can achieve better performance with fewer adversarial examples than the standard adversarial training (Standard-AT), by utilizing semantic features and semi-supervised learning. Such improvement in training efficiency is quite beneficial since the adversarial training process is time-consuming and the adversarial samples are often limited.

4. Detailed Design of the Model

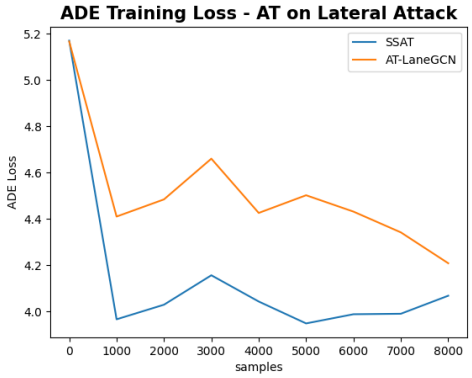
All methods' feature extractors are based on the one in [4], which is a representative graph-based method. The feature extractor consists of an actor network, a map network and four types of attention networks. The actor network uses three 1-D convolutions to extract multi-scale features from time-series input trajectory data. The map network is designed to learn the structured map representation from vectorized map input. Then, the extracted features are fed into a fusion network that calculates different attentions of actor-to-actor, actor-to-lane, lane-to-actor, and lane-to-lane. Finally, 128-dimensional features are generated for downstream tasks.

Our AAE encoder is based on a layer of linear residual block and three heads to predict the three different types of latent variables.

As mentioned in the main paper, we divide the latent space into three different parts, representing the latent semantics (intention), longitudinal semantics (time headway) and remaining information, respectively. Fig. 4 demon-



(a)



(b)

Figure 3: Evaluated by the ADE (average displacement error), we observe that our SSAT method can achieve better performance with fewer adversarial examples than the standard adversarial training method (Standard-AT). Fig. (a) is trained on the ADE attacks. Fig. (b) is trained on the lateral attacks. The experiments are conducted on the Argoverse 1 dataset.

strates the reason why we choose a log-normal distribution to regularize the longitudinal semantics.

Three discriminators with the same structure are used to regularize the distribution of latent space. They consist of two-layer fully-connected neural networks and output a distinction score after a sigmoid layer.

The decoder takes 10-dimensional latent vectors as input and projects the predicted results from the latent space to the trajectory space. The decoder is a simple three-layer fully-connected neural network.

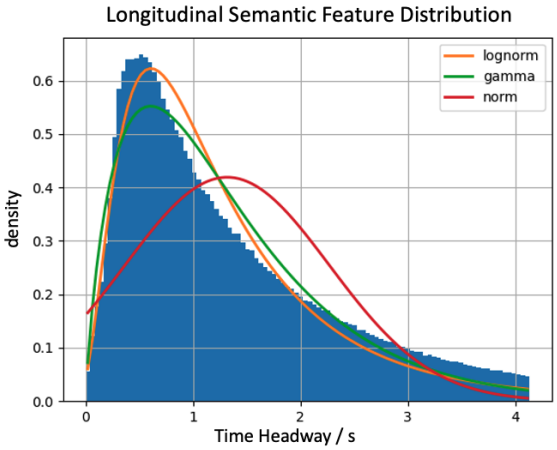


Figure 4: The histogram and fitted distributions of time headway, the longitudinal semantics our method utilizes.

5. Details on Experiment Settings

5.1. Generation of Adversarial Examples:

We use Projected Gradient Decent (PGD) [5] to generate adversarial examples for all attack types, and its parameters are set as follows in Table 2.

Table 2: PGD setting for different types of attacks.

Attacks	Max perturbation ϵ	Steps	Step size α
ADE	0.2	20	0.01
Lateral	0.3	40	0.01
Longitudinal	0.3	20	0.01
L1	0.4	20	0.01
FDE	0.4	20	0.01

5.2. SSAT Learning Rate:

For Argoverse 1 and 2, we set the learning rate for trajectory prediction loss, regularization loss, semi-supervised loss, and discrimination loss as 5e-4,1e-6,1e-5, and 1e-6, respectively. For the Apolloscape dataset that consists of simpler scenarios without map contexts, we set the learning rate for trajectory prediction loss, regularization loss, semi-supervised loss, and discrimination loss as 5e-5, 5e-7, 5e-6, and 5e-7, respectively.

6. Limitations and Future Work

Our work proposes an adversarial training method to enhance the robustness of trajectory prediction and improve its robust generalization performance on unseen attacks. However, there are some trade-offs. First, by adding a VAE architecture with disentangled latent space [6], the prediction

accuracy in benign cases will drop slightly. Then, similar to the phenomenon observed in adversarial training for other domains, there is also a trade-off between adversarial robustness and standard accuracy, as discussed in the paper. To address these challenges in future work, we plan to explore the following directions: 1) further optimizing the architecture and method for adversarial training, e.g., by adding regularization [7] and collecting more data [5], and 2) exploring adaptive system-level design. For the system-level design, we can utilize the VAE architecture to build an additional reconstruction module and detect anomalies by comparing the reconstruction error. When the input is detected as an anomaly (i.e., potential attack), the system will switch to a robust mode for trajectory prediction and also take corresponding actions in the planning module; otherwise, the system will use the original prediction mode to maintain the standard accuracy.

In addition, we also plan to study certified robustness for trajectory prediction. From the literature and our own experience in this area, there could be significant challenges for analyzing certified robustness when addressing larger neural networks and larger input disturbances, which are common in the case of adversarial attacks to trajectory prediction – the GNN and AAE networks used in our work are quite complex, and the carefully optimized adversarial examples often induce larger input disturbance at places than the small random noises typically being studied in certified robustness. One direction we are thinking of is to introduce physical dynamics to the latent space for facilitating the analysis of certified robustness.

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 1

[2] 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. 1

[3] Daniel Kang, Yi Sun, Tom Brown, Dan Hendrycks, and Jacob Steinhardt. Transfer of adversarial robustness between perturbation types. *arXiv preprint arXiv:1905.01034*, 2019. 1

[4] 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. 2

[5] 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. 3, 4

[6] Matthew Willetts, Alexander Camuto, Tom Rainforth, Stephen Roberts, and Chris Holmes. Improving vaes' robustness to adversarial attack. *arXiv preprint arXiv:1906.00230*, 2019. 3

[7] 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. 4