

# Adversarial Attacks On Multi-Agent Communication

James Tu<sup>1,2\*</sup> Tsunhsuan Wang<sup>3\*</sup> Jingkang Wang<sup>1,2</sup> Sivabalan Manivasagam<sup>1,2</sup>  
 Mengye Ren<sup>1,2</sup> Raquel Urtasun<sup>1,2</sup>

<sup>1</sup>Waabi    <sup>2</sup>University of Toronto    <sup>3</sup>MIT

{jtu, wangjk, manivasagam, mren, urtasun}@cs.toronto.edu  
 johnsonwang0810@gmail.com

## Abstract

*Growing at a fast pace, modern autonomous systems will soon be deployed at scale, opening up the possibility for cooperative multi-agent systems. Sharing information and distributing workloads allow autonomous agents to better perform tasks and increase computation efficiency. However, shared information can be modified to execute adversarial attacks on deep learning models that are widely employed in modern systems. Thus, we aim to study the robustness of such systems and focus on exploring adversarial attacks in a novel multi-agent setting where communication is done through sharing learned intermediate representations of neural networks. We observe that an indistinguishable adversarial message can severely degrade performance, but becomes weaker as the number of benign agents increases. Furthermore, we show that black-box transfer attacks are more difficult in this setting when compared to directly perturbing the inputs, as it is necessary to align the distribution of learned representations with domain adaptation. Our work studies robustness at the neural network level to contribute an additional layer of fault tolerance to modern security protocols for more secure multi-agent systems.*

## 1. Introduction

With rapid improvements of modern autonomous systems, it is only a matter of time until they are deployed at scale, opening up the possibility of cooperative multi-agent systems. Individual agents can benefit greatly from shared information to better perform their tasks [26, 59]. For example, by aggregating sensory information from multiple viewpoints, a fleet of vehicles can perceive the world more clearly, providing significant safety benefits [52]. Moreover, in a network of connected devices, distributed processing across multiple agents can improve computation ef-

ficiency [18]. While cooperative multi-agent systems are promising, relying on communication between agents can pose security threats as shared information can be malicious or unreliable [54, 3, 37].

Meanwhile, modern autonomous systems typically rely on deep neural networks known to be vulnerable to adversarial attacks. Such attacks craft small and imperceptible perturbations to drastically change a neural network’s behavior and induce false outputs [48, 21, 8, 30]. Even if an attacker has the freedom to send any message, such small perturbations may be the most dangerous as they are indistinguishable from their benign counterparts, making corrupted messages difficult to detect while still highly malicious.

While modern cyber security algorithms provide adequate protection against communication breaches, adversarial robustness of multi-agent deep learning models has yet to be studied. Meanwhile, when it comes to safety-critical applications like self-driving, additional layers of redundancy and improved security are always welcome. Thus, by studying adversarial robustness, we can enhance modern security protocols by introducing an additional layer of fault tolerance at the neural network level.

Adversarial attacks have been studied extensively but existing approaches mostly consider attacks on input domains like images [48, 21], point clouds [7, 50], and text [44, 14]. On the other hand, multi-agent systems often distribute computation across different devices and transmit intermediate representations instead of input sensory information [52, 18]. Specifically, when deep learning inference is distributed across different devices, agents will communicate by transmitting feature maps, which are activations of intermediate neural network layers. Such learned communication has been shown to be superior due to transmitting compact but expressive messages [52] as well as efficiently distributing computation [18].

In this paper, we investigate adversarial attacks in this novel multi-agent setting where perturbations are applied to learned intermediate representations. An illustration is

\*Equal contribution.

Work done while all authors were at UberATG.

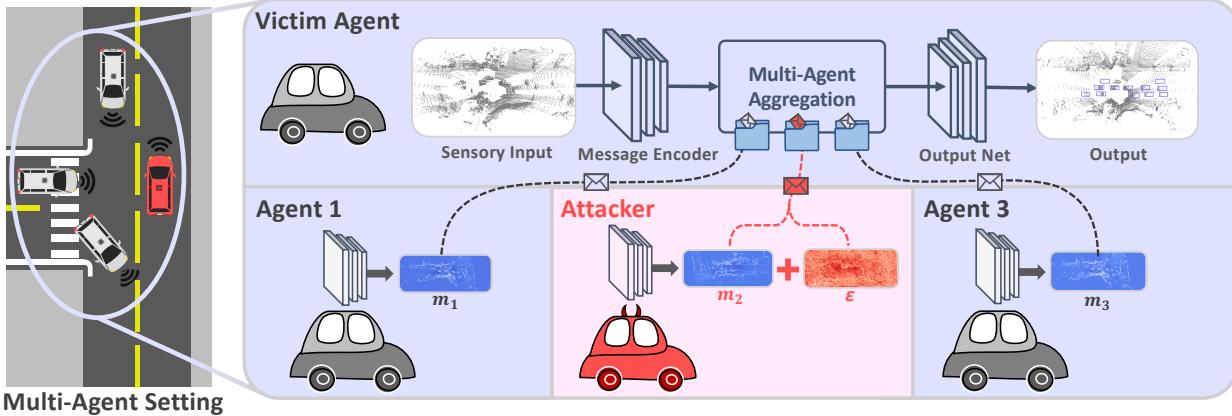


Figure 1. Overview of a multi-agent setting with one malicious agent (red). Here the malicious agent attempts to sabotage a victim agent by sending an adversarial message. The adversarial message is indistinguishable from the original, making the attack difficult to detect.

shown in Figure 1. We conduct experiments and showcase vulnerabilities in two highly practical settings: multi-view perception from images in a fleet of drones and multi-view perception from LiDAR in a fleet of self-driving vehicles (SDVs). By leveraging information from multiple viewpoints, these multi-agent systems are able to significantly outperform those that do not exploit communication.

We show, however, that perturbed transmissions which are indistinguishable from the original can severely degrade the performance of receivers particularly as the ratio of malicious to benign agents increases. With only a single attacker, as the number of benign agents increase, attacks become significantly weaker as aggregating more messages decreases the influence of malicious messages. When multiple attackers are present, they can coordinate and jointly optimize their perturbations to strengthen the attack. In terms of defense, when the threat model is known, adversarial training is highly effective, and adversarially trained models can defend against perturbations almost perfectly and even slightly enhance performance on natural examples. Without knowledge of the threat model, we can still achieve reasonable adversarial robustness by designing more robust message aggregation modules.

We then move on to more practical attacks in a black box setting where the model is unknown to the adversary. Since query-based black box attacks need to excessively query a target model that is often unaccessible, we focus on query-free transfer attacks that are more feasible in practice. However, transfer attacks are much more difficult to execute at the feature-level than on input domains. In particular, since perturbation domains are model dependent, vanilla transfer attacks are ineffective because two neural networks with the same functionality can have very different intermediate representations. Here, we find that training the surrogate model with domain adaptation is key to aligning the distribution of intermediate features and achieve much better transferabil-

ity. To further enhance the practicality of attacks, we propose to exploit the temporal consistency of sensory information processed by modern autonomous systems. When frames of sensory information are collected milliseconds apart, we can exploit the redundancy in adjacent frames to create efficient, low-budget attacks in an online manner.

## 2. Related Work

**Multi-Agent Deep Learning Systems:** Multi-agent and distributed systems are widely employed in real-world applications to improve computation efficiency [27, 17, 2], collaboration [52, 59, 18, 41, 42], and safety [38, 35]. Recently, autonomous systems have improved greatly with the help of neural networks. New directions have opened up in cooperative multi-agent deep learning systems e.g., federated learning [27, 2]. Although multi-agent communication introduces a multitude of benefits, communication channels are vulnerable to security breaches, as communication channels can be attacked [34, 45], encryption algorithms can be broken [46], and agents can be compromised [5, 61]. Thus, imperfect communication channels may be used to execute adversarial attacks which are especially effective against deep learning systems. While robustness has been studied in the context of federated learning [20, 1, 56, 19], the threat models are different as dataset poisoning and model poisoning are typically used. To the best of our knowledge, few works study adversarial robustness on multi-agent deep learning systems during inference.

**Adversarial Attacks:** Adversarial attacks were first discovered in the context of image classification [48], where a small imperceptible perturbation can drastically change a neural network’s behaviour and induce false outputs. Such attacks were then extended to various applications such as semantic segmentation [57] and reinforcement learning [24]. There are two main settings for adversarial at-

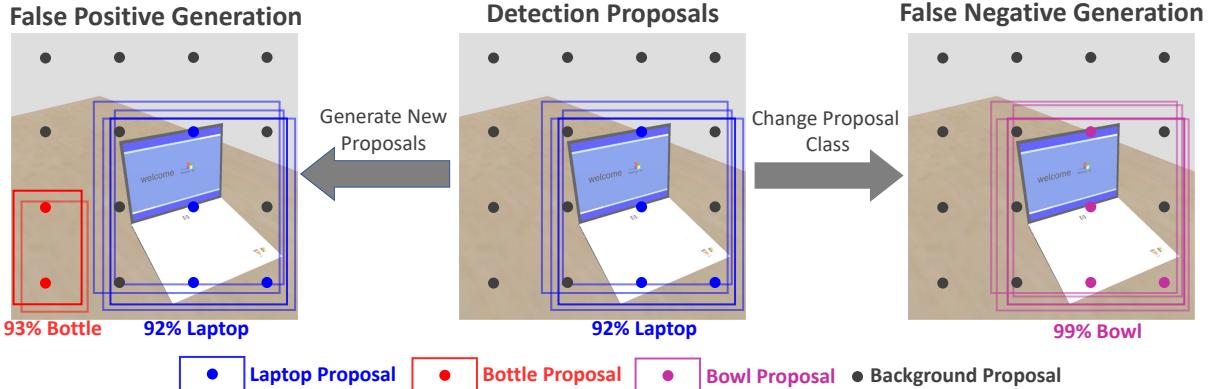


Figure 2. **Attacking object detection proposals:** False positives are created by changing the class of background proposals and false negatives are created by changing the class of the original proposals.

tacks - *white box* and *black box*. In a white box setting [48, 21, 30], the attacker has full access to the target neural network weights and adversarial examples can be generated using gradient-based optimization to maximize the network’s error. In contrast, black box attacks are conducted without knowledge of the target neural network weights and therefore without any gradient computation. In this case, attackers can leverage real world knowledge to inject adversaries that resemble common real world objects [47, 36]. However, if the attacker is able to query the target model, the literature proposes several different strategies to perform query-based attacks [4, 12, 6, 10]. However, query-based attacks are infeasible for some applications as they typically require prohibitively large amounts of queries and computation. Apart from query-based attacks, a more practical but more challenging alternative is to conduct transfer attacks [39, 58, 16] which do not require querying the target model. In this setting, the attacker trains a surrogate model that imitates the target model. By doing so, the hope is that perturbations generated for the surrogate model will transfer to the target model.

**Perturbations In Feature Space:** While most works in the literature focus on input domains like images, some prior works have considered perturbations on intermediate representations within neural networks. Specifically, [25] estimated the projection of adversarial gradients on a selected subspace to reduce the queries to a target model. [40, 44, 14] proposed to generate adversarial perturbation in word embeddings for finding adversarial but semantically-close substitution words. [55, 60] showed that training on adversarial embeddings could improve the robustness of Transformer-based models for NLP tasks.

### 3. Attacks On Multi-Agent Communication

This section first introduces the multi-agent framework in which agents leverage information from multiple view-

points by transmitting intermediate feature maps. We then present our method for generating adversarial perturbations in this setting. Moving on to more practical settings, we consider black box transfer attacks and find that it is necessary to align the distribution of intermediate representations. Here, training a surrogate model with domain adaptation can create transferable perturbations. Finally, we show efficient online attacks by exploiting the temporal consistency of sensory inputs collected at high frequency.

#### 3.1. Multi-Agent Communication

We consider a setting where multiple agents cooperate to better perform their tasks by sharing observations from different viewpoints encoded via a learned intermediate representation. Adopting prior work [52], we assume a homogeneous set of agents using the same neural network. Then, each agent  $i$  processes sensor input  $x_i$  to obtain an intermediate representation  $m_i = F(x_i)$ . The intermediate feature map is then broadcasted to other agents in the scene. Upon receiving messages, agent  $j$  will aggregate and process all incoming messages to generate output  $Z_j = G(m_1, \dots, m_N)$ , where  $N$  is the number of agents. Suppose that an attacker agent  $i$  targets a victim agent  $j$ . Here, the attacker attempts to send an indistinguishable adversarial message  $m'_i = m_i + \delta$  to maximize the error in  $Z'_j = G(m_1, \dots, m_i + \delta, m_N)$ . The perturbation  $\delta$  is constrained by  $\|\delta\|_p \leq \epsilon$  to ensure that the malicious message is subtle and difficult to detect. An overview of the multi-agent setting is shown in Figure 1.

In this paper, we specifically focus on object detection as it is a challenging task where aggregating information from multiple viewpoints is particularly helpful. In addition, many downstream robotics tasks depend on detection and thus a strong attack can jeopardize the performance of the full system. In this case, output  $Z$  is a set of  $M$  bounding box proposals  $z^{(1)}, \dots, z^{(M)}$  at different spatial locations. Each proposal consists of class scores  $z_{\sigma_0}, \dots, z_{\sigma_k}$

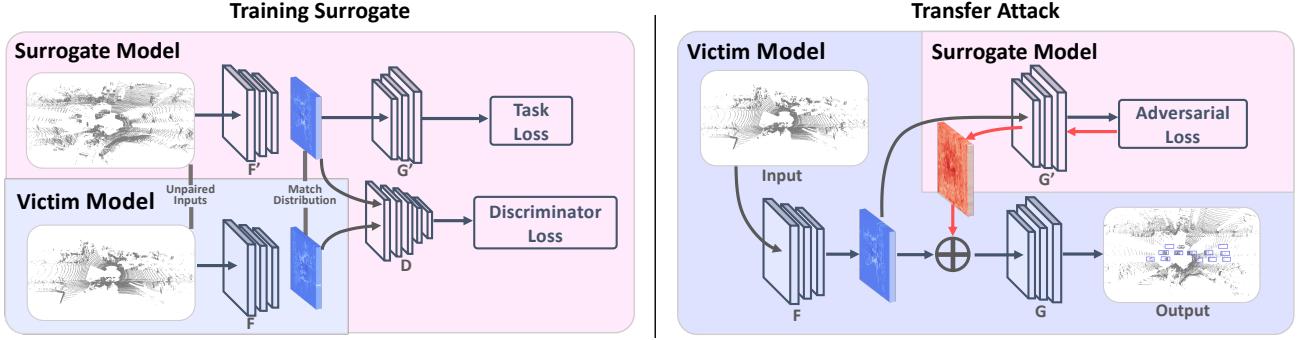


Figure 3. Our proposed transfer attack which incorporates domain adaptation when training the surrogate model. During training, the discriminator forces  $F'$  to produce intermediate representations similar to  $F$ . As a result,  $G'$  can generate perturbations that transfer to  $G$ .

and bounding box parameters describing the spatial location and dimensions of the bounding box. Here classes  $0, \dots, k - 1$  are the object classes and  $k$  denotes the background class where no objects are detected.

When performing detection, models try to output the correct object class  $k$  and maximize the ratio of intersection over union (IOU) of the proposed and ground truth bounding boxes. In a post processing step, proposals with high confidence are selected and overlapping bounding boxes are filtered with non-maximum suppression (NMS) to ideally produce a single estimate per ground truth object.

### 3.2. Adversarial Perturbation Generation

We first introduce our loss objective for generating adversarial perturbations against object detection. To generate false outputs, we aim to confuse the proposal class. For detected objects, we suppress the score of the correct class to generate false negatives. For background classes, false positives are created by pushing up the score of an object class. In addition, we also aim to minimize the intersection-over-union (IoU) of the bounding box proposals to further degrade performance by producing poorly localized objects. We define the adversarial loss of the perturbed output  $z'$  with respect to an unperturbed output  $z$  instead of the ground truth, as it may not always be available to the attacker. For each proposal  $z$ , let  $u = \text{argmax}_i\{z_{\sigma_i} | i = 0 \dots m\}$  be the highest confidence class. Given the original object proposal  $z$  and the proposal after perturbation  $z'$ , our loss function tries to push  $z'$  away from  $z$ :

$$\ell_{adv}(z', z) = \begin{cases} -\log(1 - z'_{\sigma_u}) \cdot \text{IoU}(z', z) & \text{if } u \neq k \text{ and } z_{\sigma_u} > \tau^+, \\ -\lambda \cdot z'_{\sigma_v} \gamma \log(1 - z'_{\sigma_v}) & \text{if } u = k \text{ and } z_{\sigma_u} > \tau^-, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

An illustration of the attack objective is shown in Figure 2. When  $u \neq k$  and the original prediction is not a background class, we apply an untargetted loss to reduce the likelihood of the intended class. When the intended pre-

diction is the background class  $k$ , we specifically target a non-background class  $v$  to generate a false positive. We simply choose  $v$  to be the class with the highest confidence that is not the background class. The IoU operator denotes the intersection over union of two proposals,  $\lambda$  is a weighting coefficient, and  $\tau^-, \tau^+$  filter out proposals that are not confident enough. We provide more analysis and ablations to justify our loss function design in our experiments.

Following prior work [50], it is necessary to minimize the adversarial loss over all proposals. Thus, the optimal perturbation under an  $\epsilon - \ell_p$  bound is

$$\delta^* = \underset{\|\delta\|_p \leq \epsilon}{\operatorname{argmin}} \sum_{m=1}^M \ell_{adv}(z'^{(m)}, z^{(m)}). \quad (2)$$

Our work considers an infinity norm  $p = \infty$  and we minimize this loss across all proposals using projected gradient descent (PGD) [31], clipping  $\delta$  to be within  $[-\epsilon, \epsilon]$ .

### 3.3. Transfer Attack

We also consider transfer attacks as they are the most practical. White box attacks assume access to the victim model's weights which is difficult to obtain in practice. On the other hand, query-based optimization is too expensive to execute in real time as state-of-the-art methods still require thousands of queries [13, 11] on CIFAR-10. Instead, when we do not have access to the weights of the victim model  $G$ , we can imitate it with a surrogate model  $G'$  such that perturbations generated by the surrogate model can transfer to the target model.

One major challenge for transfer attacks in our setting is that perturbations are generated on intermediate feature maps. Our experiments show that vanilla transfer attacks are almost completely ineffective as two networks with the same functionality do not necessarily have the same intermediate representations. When training  $F$  and  $G$ , there is no direct supervision on the intermediate features  $m = F(x)$ . Therefore, even with the same architecture, dataset,

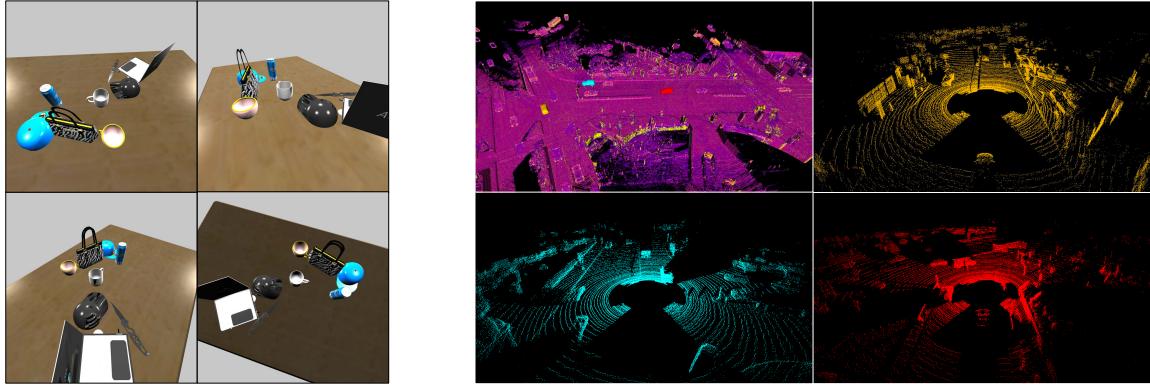


Figure 4. Two multi-agent datasets we use. On the left are images of ShapeNet objects taken from different view points. On the right are LiDAR sweeps by different vehicles in the same scene.

and training schedule, a surrogate  $F'$  may produce messages  $m'$  with very different distribution from  $m$ . As an example, a permutation of feature channels carries the same information but results in a different distribution. In general, different random seeds, network initializations or non-deterministic GPU operations can result in different intermediate representations. It follows that if  $m'$  does not faithfully replicate  $m$ , we cannot expect  $G'$  to imitate  $G$ .

Thus, to execute transfer attacks, we must have access to samples of the intermediate feature maps. Specifically, we consider a scenario where the attacker can spy on the victim’s communication channel to obtain transmitted messages. However, since sensory information is not transmitted, the attacker does not have access to pairs of input  $x$  and intermediate representation  $m$  to directly supervise the surrogate  $F'$  via distillation. Thus, we propose to use Adversarial Discriminative Domain Adaptation (ADDA) [51] to align the distribution of  $m$  and  $m'$  without explicit input-feature pairs. An overview is shown in Figure 3.

In the original training pipeline,  $F'$  and  $G'$  would be trained to minimize task loss

$$\mathcal{L}_{task}(z, y, b) = \begin{cases} -\log(z_{\sigma_y}) - \text{IoU}(z, b) & \text{if } y \neq k, \\ -\log(z_{\sigma_y}) & \text{if } y = k, \end{cases} \quad (3)$$

where  $b$  is a ground truth bounding box and  $y$  is its class. The task loss maximizes the log likelihood of the correct class and the IoU between the proposal box and the ground truth box. In addition, we encourage domain adaptation by introducing a discriminator  $D$  to distinguish between real messages  $m$  and surrogate messages  $m'$ . The three modules  $F'$ ,  $G'$ , and  $D$  can be optimized using the following min-max criterion:

$$\min_{F'} \max_{G'} \mathcal{L}_{task}(x) + \beta [\log D(F(x)) + \log(1 - D(F'(x)))] \quad (4)$$

where  $\beta$  is a weighting coefficient and we use binary cross entropy loss to supervise the discriminator. During training,

we adopt spectral normalization [33] in the discriminator and the two-time update rule [22] for stability.

### 3.4. Online Attack

In modern applications of autonomous systems, consecutive frames of sensory information are typically collected only milliseconds apart. Thus, there is a large amount of redundancy between consecutive frames which can be exploited to achieve more efficient adversarial attacks. Following previous work [53] in images, we propose to exploit this redundancy by using the perturbation from the previous time step as initialization for the current time step.

Furthermore, we note that intermediate feature maps capture the spatial context of sensory observations, which change due to the agent’s egomotion. Therefore, by applying a rigid transformation on the perturbation at every time step to account for egomotion, we can generate stronger perturbations that are synchronized with the movement of sensory observations relative to the agent. In this case, the perturbations are updated as follows:

$$\delta^{(t+1)} \leftarrow H_{t \rightarrow t+1}(\delta^{(t)}) - \alpha \nabla_{H_{t \rightarrow t+1}(\delta)} \mathcal{L}_{adv}(Z'^{(t+1)}, Z^{(t+1)}). \quad (5)$$

Here  $H_{t \rightarrow t+1}$  is a rigid transformation mapping the attacker’s pose at time  $t$  to  $t+1$  and  $\alpha$  is the step size. By leveraging temporal consistency we can generate strong perturbations with only one gradient update per time step, making online attacks more feasible.

## 4. Experiments

### 4.1. Multi-Agent Settings

**Multi-View ShapeNet:** We conduct our attacks on multi-view detection from images, which is a common task for a fleets of drones. Following prior work [15], we generate a synthetic dataset by placing 10 classes of ShapeNet [9] objects on a table (see Figure 4). From each class, we sub-

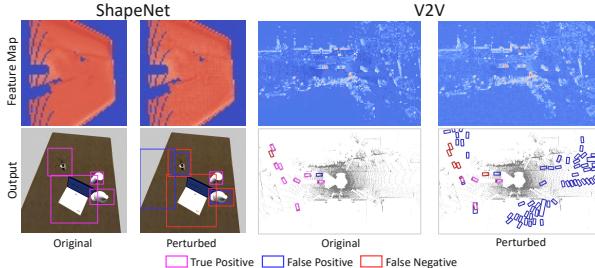


Figure 5. Qualitative attack examples. Top: Messages sent by another agent visualized in bird’s eye view. Bottom: outputs. Perturbations are very subtle but severely degrade performance.

sample 50 meshes and use a 40/10 split for training and validation. In every scene, we place 4 to 8 objects and perform collision checking to ensure objects do not overlap. Then, we capture  $128 \times 128$  RGB-D images from 2 to 7 viewpoints sampled from the upper half of a sphere centered at the table center with a radius of 2.0 units. This dataset consists of 50,000 training scenes and 10,000 validation scenes. When conducting attacks, we randomly sample one of the agents to be the adversary. Our detection model uses an architecture similar to the one introduced in [15]. Specifically, we process input RGB-D images using a U-Net [43] and then unproject the features into 3D using the depth measures. Features from all agents are then warped into the same coordinate frame and aggregated with mean pooling. Finally, aggregated features are processed by a 3D U-Net and a detection header to generate 3D bounding box proposals.

**Vehicle To Vehicle Communication:** We also consider a self-driving setting with vehicle-to-vehicle(V2V) communication. Here, we adopt the dataset used in [52], where 3D reconstructions of logs of real world LiDAR scans are simulated from the perspectives of other vehicles in the scene using a high-fidelity LiDAR simulator [32]. These logs are collected by self-driving vehicles equipped with LiDAR sensors capturing 10 frames per second (see Figure 4). The training set consists of 46,796 subsampled frames from the logs and we do not subsample the validation set, resulting in 96,862 frames. In every log we select one attacker vehicle and sample others to be cooperative agents with up to 7 agents in each frame unless otherwise specified. This results in a consistent assignment of attackers and V2V agents throughout the frames. In this setting, we use the state-of-the-art perception and motion forecasting model V2VNet [52]. Here, LiDAR inputs are first encoded into bird’s eye view (BEV) feature maps. Feature maps from all agents are then warped into the ego coordinate frame and aggregated with a GNN to produce BEV bounding box proposals. More details of the ShapeNet model and V2VNet are provided in the supplementary material.

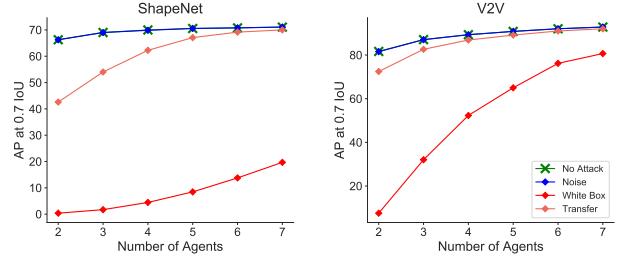


Figure 6. Evaluation under no perturbation, uniform noise, transfer attack, and white box attack. Results are grouped by the number of agents in the scene where one agent is the attacker.

	ShapeNet		V2V	
	Clean	Perturbed	Clean	Perturbed
Original	66.33	0.62	82.19	7.55
Adv Trained	<b>67.29</b>	<b>66.00</b>	<b>82.60</b>	<b>83.44</b>

Table 1. Results of adversarial training. Robustness increases significantly, matching clean inference. Furthermore performance on clean data also improves slightly.

**Implementation Details:** When conducting attacks, we set  $\epsilon = 0.1$ . For the proposed loss function, we set  $\lambda = 0.2$ ,  $\tau^- = 0.7$ ,  $\tau^+ = 0.3$ , and  $\gamma = 1$ . Projected gradient descent is done using Adam with learning rate 0.1 and we apply 15 PGD steps for ShapeNet and only 1 PGD step for low budget online attacks in the V2V setting. The surrogate models use the same architecture and dataset as the victim models. When training the surrogate model, we set  $\beta = 0.01$ , model learning rate 0.001, and discriminator learning rate 0.0005. For evaluation, we compute area under the precision-recall curve of bounding boxes, where bounding boxes are correct if they have an IoU greater than 0.7 with a ground truth box of the same class. We refer to this metric as *AP at 0.7* in the following.

## 4.2. Results

**Attack Results:** Visualizations of our attack are shown in Figure 5 and we present quantitative results of our attack and baselines in Figure 6. We split up the evaluation by the number of agents in the scene and one of the agents is always an attacker. As a baseline, we sample the perturbation from  $\mathcal{U}(-\epsilon, \epsilon)$  to demonstrate that the same  $\epsilon$  bounded uniform noise does not have any impact on detection performance. The white box attack is especially strong when few agents are in the scene, but becomes weaker as the number of benign agents increase, causing the relative weight of the adversarial features in mean pooling layers to decrease. Finally, our transfer attack with domain adaptation achieves moderate success with few agents in the scene, but is significantly weaker than the white box attack.

Agents	Clean			Perturbed		
	2	4	6	2	4	6
Mean Pool	82.09	89.25	92.43	0.90	12.93	41.77
GNN(Mean)	<b>82.19</b>	<b>89.93</b>	<b>92.94</b>	7.55	52.31	76.18
GNN(Median)	82.11	87.12	90.75	12.8	<b>67.70</b>	<b>86.30</b>
GNN(Soft Med)	82.19	89.67	92.49	<b>21.53</b>	61.37	84.99

Table 2. Choice of fusion in V2VNet affects performance and robustness. We investigate using mean pooling and using a GNN with various aggregation methods.

**Robustifying Models:** To defend against our proposed attack, we conduct adversarial training against the white box adversary and show the results in Table 1. Here, we follow the standard adversarial training set up, except perturbations are applied to intermediate features instead of inputs. This objective can be formulated as

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} \max_{\|\delta\|_{\infty} < \epsilon} \phi((x, y, \delta); \theta) := \mathcal{L}_{task}(G(F(x_0), \dots, F(x_i) + \delta, \dots, F(x_N); \theta)), \quad (6)$$

where  $D$  is the natural training distribution and  $\theta$  denotes model parameters. During training, we generate a new perturbation  $\delta$  for each training sample. In the multi-agent setting, we find it easier to recover from adversarial perturbations when compared to traditional single-agent attacks. Moreover, adversarial training is able to slightly improve performance on clean data as well, while adversarial training has been known to hurt natural performance in previous settings [28, 49].

While adversarial training is effective in this setting, it requires knowledge of the threat model. When the threat model is unknown, we can still naturally boost the robustness of multi-agent models with the design of the aggregation module. Specifically, we consider several alternatives to V2VNet’s GNN fusion and present the performance under attacked and clean data in Table 2. First, replacing the entire GNN with an adaptive mean pooling layer significantly decreases robustness. On the other hand, we swap out the mean pooling in GNN nodes with median pooling and find that it increases robustness at the cost of performance on clean data with more agents, since more information is discarded. We refer readers to the supplementary materials for more details on implementation of the soft median pooling.

**Multiple Attackers:** We previously focused on settings with one attacker, and now conduct experiments with multiple attackers in the V2V setting. In each case, we also consider if attackers are able to cooperate. In cooperation, attackers jointly optimize their perturbations. Without cooperation, attackers are blind to each other and optimize

Agents	Cooperative			Non-Cooperative		
	4	5	6	4	5	6
1 Attacker	52.31	65.00	76.18	52.31	65.00	76.18
2 Attacker	28.31	41.34	54.50	39.02	51.96	64.02
3 Attacker	<b>12.07</b>	<b>22.84</b>	<b>35.13</b>	<b>24.27</b>	<b>38.17</b>	<b>51.58</b>

Table 3. Multiple white box attackers in the V2V setting. Cooperative attackers jointly optimize their perturbations and non-cooperative attackers optimize without knowledge of each other.

Attackers	0	1	2	3
Train On 0	89.93	52.31	28.31	12.07
Train On 1	<b>90.09</b>	<b>90.00</b>	81.95	75.28
Train On 2	89.71	89.68	88.91	88.33
Train On 3	89.55	89.51	<b>88.94</b>	<b>88.51</b>

Table 4. Adversarial training with multiple attackers in the V2V setting. We train on settings with various number of attackers and evaluate the models across the settings.

their perturbations assuming other messages have not been perturbed. Results with up to 3 attackers are shown in Table 3. As expected, more attackers can increase the strength of attack significantly, furthermore, if multiple agents can coordinate, a stronger attack can be generated.

Next, we apply adversarial training to the multi-attacker setting and present results in Table 4. Here, all attacks are done in the cooperative setting and we show results with 4 total agents. Similar to the single attacker setting, adversarial training is highly effective. However, while adversarial training against one attacker improves performance in natural examples, being robust to stronger attacks sacrifices performance on natural examples. This suggests that adversarial training has the potential to improve general performance when an appropriate threat model is selected. Furthermore, we can see that training on fewer attacks does not generalize perfectly to more attackers but the opposite is true. Thus, it is necessary to train against an equal or greater threat model to fully defend against such attacks.

**Domain Adaptation:** More results of the transfer attack are included in Table 5. First, we conduct an ablation and show that a transfer attack without domain adaptation (DA) is almost completely ineffective. On the contrary, surrogate models trained with DA achieve significant improvements. A visual demonstration of feature map alignment with DA is shown in Figure 7, visualizing 4 channels of the intermediate feature maps. Features from a surrogate trained with DA are visually very similar to the victim, while a surrogate trained without DA produces features with no resemblance.

Since our proposed DA improves the transferability of the surrogate model, we can further improve our transfer

	ShapeNet	V2V
Clean	66.28	82.19
Transfer	66.21	81.31
Transfer + DA	42.59	72.45
Transfer + DA + ILAP	<b>35.69</b>	<b>71.76</b>
Transfer + DA + DI	49.38	75.18

Table 5. Transfer attacks evaluated with 2 agents. Training the surrogate with domain adaptation (DA) significantly improves transferability. In addition, we attempt to enhance transferability with ILAP [23] and DI [58].

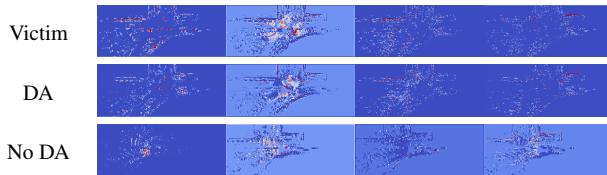


Figure 7. Visualization of how domain adaptation(DA) affects 4 channels of the intermediate feature map. Observe that the surrogate trained with DA closely imitates the victim model, while the surrogate trained without DA produces different features.

attack by also adopting methods from the literature which enhance the transferability of a given perturbation. We find that generating perturbations from diversified inputs (DI) [58] is ineffective as resizing input feature maps distorts spatial information which is important for localizing objects detection. On the other hand, using an intermediate level attack projection (ILAP) [23] yields a small improvement. Overall, we find transfer attacks more challenging when at the feature level. In standard attacks on sensory inputs, perturbations are transferred into the same input domain. However, at a feature level the input domains are model-dependent, making transfer attacks between different models more difficult.

**Online Attacks:** We conduct an ablation on the proposed methods for exploiting temporal redundancy in an online V2V setting, shown in Table 6. First, if we ignore temporal redundancy and do not reuse the previous perturbation, attacks are much weaker. In this evaluation we switch from PGD [31] to FGSM [21] to obtain a stronger perturbation in one update for fair comparison. We also show that applying a rigid transformation on the perturbations at every frame to compensate for egomotion provides a modest improvement to the attack when compared to the *No Warp* ablation.

**Loss Function Design:** We conduct an ablation study on using our adversarial loss  $\mathcal{L}_{adv}$  instead of the negative task loss  $-\mathcal{L}_{task}$  in Table 7. This ablation validates our loss function and showcase that for structured outputs, properly designed adversarial losses is more effective than the naive negative task loss which is widely

	2 Agents	4 Agents	6 Agents
Our Attack	7.55	<b>52.31</b>	<b>76.18</b>
No Warping	<b>7.17</b>	52.35	77.37
Independent	56.98	80.21	87.05

Table 6. Ablation on online attacks in the V2V setting. *Independent* refers to treating each frame independently and not reusing previous perturbations. *No warp* refers to omitting the rigid transformation to account for egomotion.

	2 Agents	4 Agents	6 Agents
ShapeNet	$-\mathcal{L}_{task}$	6.10	20.07
	$\mathcal{L}_{adv}$	<b>0.37</b>	<b>4.45</b>
V2V	$-\mathcal{L}_{task}$	20.8	63.82
	$\mathcal{L}_{adv}$	<b>7.55</b>	<b>52.31</b>
			<b>76.18</b>

Table 7. Ablation on loss function, it produces stronger adversarial attacks than simply using the negative of the training task loss.

used in image classification tasks. Our choice for the loss function design is motivated by our knowledge of the post-processing non-maximum suppression (NMS). Since NMS selects bounding boxes with the highest confidence in a local region, proposals with higher scores should receive stronger gradients. More specifically, an appropriate loss function of  $f$  for proposal score  $\sigma$  should satisfy  $(|\nabla_{\sigma_2} f(\sigma_2)| - |\nabla_{\sigma_1} f(\sigma_1)|) / (\sigma_2 - \sigma_1) > 0$  so that  $|\nabla_{\sigma} f(\sigma)|$  is monotonically increasing in  $\sigma$ . We can see that the standard log likelihood does not satisfy this criteria, which explains why our loss formulation is more effective. In addition, we add the focal loss term [29] to generate more false positives, as aggressively focusing on one proposal in a local region is more effective due to NMS.

## 5. Conclusion

In this paper, we investigate adversarial attacks on communication in multi-agent deep learning systems. Our experiments in two practical settings demonstrate that compromised communication channels can be used to execute adversarial attacks. However, robustness increases as the ratio of benign to malicious actors increases. Furthermore, we found that more practical transfer attacks are more challenging in this setting and require aligning the distributions of intermediate representations. Finally, we propose a method to achieve efficient and practical online attacks by exploiting temporal consistency of sensory inputs. We believe studying adversarial robustness on multi-agent deep learning models in real-world applications is an important step towards more secure multi-agent systems.

## References

- [1] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin B. Calo. Analyzing federated learning through an adversarial lens. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 634–643. PMLR, 2019. 2
- [2] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konecný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roslander. Towards federated learning at scale: System design. *CoRR*, 2019. 2
- [3] Niklas Borselius. Mobile agent security. *Electronics & Communication Engineering Journal*, 14(5), 2002. 1
- [4] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018. 3
- [5] Thomas Brewster. Watch chinese hackers control tesla’s brakes from 12 miles away, 2016. 2
- [6] Thomas Brunner, Frederik Diehl, Michael Truong-Le, and Alois Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. *CoRR*, abs/1812.09803, 2018. 3
- [7] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z. Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *CCS*, 2019. 1
- [8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *SP*, 2017. 1
- [9] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. *arXiv preprint arXiv:1512.03012*, 2015. 5
- [10] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. *arXiv preprint arXiv:1904.02144*, 3, 2019. 3
- [11] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy*, pages 1277–1294. IEEE, 2020. 4
- [12] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec*, 2017. 3
- [13] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jin-Feng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*, 2019. 4
- [14] Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, and Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *CoRR*, abs/1803.01128, 2018. 1, 3
- [15] Ricson Cheng, Ziyan Wang, and Katerina Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. In *NeurIPS*. 2018. 5, 6
- [16] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *NeurIPS*, 2019. 3
- [17] Tharam S. Dillon, Chen Wu, and Elizabeth Chang. Cloud computing: Issues and challenges. In *AINA*, 2010. 2
- [18] Amir Erfan Eshratifar and Massoud Pedram. Energy and performance efficient computation offloading for deep neural networks in a mobile cloud computing environment. In *GLSVLSI*, 2018. 1, 2
- [19] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX Security Symposium*, pages 1605–1622. USENIX Association, 2020. 2
- [20] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019. 2
- [21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015. 1, 3, 8
- [22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 5
- [23] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge J. Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. *CoRR*, abs/1907.10823, 2019. 8
- [24] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. 2017. 2
- [25] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *ACM MM*, 2019. 3
- [26] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492, 2016. 1
- [27] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *CoRR*, 2016. 2
- [28] Saehyun Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *CVPR*, pages 269–278. IEEE, 2020. 7
- [29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 8
- [30] 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. 1, 3
- [31] 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. 4, 8
- [32] Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyan Zeng, Wei-Chiu Ma, Mikita Sazanovich, Bin

- Yang, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *CVPR*, 2020. 6
- [33] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 5
- [34] Bassem Mokhtar and Mohamed Azab. Survey on security issues in vehicular ad hoc networks. *Alexandria engineering journal*, 54(4):1115–1126, 2015. 2
- [35] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Technical report, 2019. 2
- [36] Ben Nassi, Dudi Nassi, Raz Ben-Netanel, Yisroel Mirsky, Oleg Drokin, and Yuval Elovici. Phantom of theadas: Phantom attacks on driver-assistance systems. *IACR*, 2020:85, 2020. 3
- [37] Petr Novák, Milan Rollo, Jirí Hodík, and Tomás Vlcek. Communication security in multi-agent systems. In *CEEMAS*, volume 2691 of *Lecture Notes in Computer Science*. Springer, 2003. 1
- [38] Marcus Obst, Laurens Hobert, and Pierre Reisdorf. Multi-sensor data fusion for checking plausibility of v2v communications by vision-based multiple-object tracking. In *VNC*, 2014. 2
- [39] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *AsiaCCS*, 2017. 3
- [40] Nicolas Papernot, Patrick D. McDaniel, Ananthram Swami, and Richard E. Harang. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM*, 2016. 3
- [41] Andreas Rauch, Felix Klanner, Ralph Rasshofer, and Klaus Dietmayer. Car2x-based perception in a high-level fusion architecture for cooperative perception systems. In *IV*, 2012. 2
- [42] Matthias Rockl, Thomas Strang, and Matthias Kranz. V2v communications in automotive multi-sensor multi-target tracking. In *VTC*, 2008. 2
- [43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6
- [44] Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. Interpretable adversarial perturbation in input embedding space for text. In *IJCAI*, 2018. 1, 3
- [45] Hichem Sedjelmaci and Sidi Mohammed Senouci. An accurate and efficient collaborative intrusion detection framework to secure vehicular networks. *Computers & Electrical Engineering*, 43, 2015. 2
- [46] Catherine Stupp and James Rundle. Capital one breach highlights shortfalls of encryption, 2019. 2
- [47] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX Security*, pages 877–894, 2020. 3
- [48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*, 2014. 1, 2, 3
- [49] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR (Poster)*. OpenReview.net, 2019. 7
- [50] James Tu, Mengye Ren, Sivabalan Manivasagam, Min Liang, Bin Yang, Richard Du, Cheng Frank, and Raquel Urtasun. Towards physically realistic adversarial examples for lidar object detection. *arXiv*, 2020. 1, 4
- [51] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 5
- [52] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. *arXiv*, 2020. 1, 2, 3, 6
- [53] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *AAAI*, 2019. 5
- [54] H. Chi Wong and Katia P. Sycara. Adding security and trust to multiagent systems. *Applied Artificial Intelligence*, 14(9):927–941, 2000. 1
- [55] Yi Wu, David Bamman, and Stuart J. Russell. Adversarial training for relation extraction. In *EMNLP*, 2017. 3
- [56] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: distributed backdoor attacks against federated learning. In *ICLR*. OpenReview.net, 2020. 2
- [57] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017. 2
- [58] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 3, 8
- [59] Tengchan Zeng, Mohammad Mozaffari, Omid Semiari, Walid Saad, Mehdi Bennis, and Merouane Debbah. Wireless communications and control for swarms of cellular-connected uavs. In *ACSSC*, 2018. 1, 2
- [60] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*, 2020. 3
- [61] Zeljka Zorz. Researchers hack bmw cars, discover 14 vulnerabilities, 2018. 2