Multi-modal Semantic Segmentation Defenses against Point Cloud Attacks in Unsupervised Domain Adaptation

Xiaoming, Xiaoqiang

***Abstract*—Opps!**

***Index Terms*—Domain adaptation, point cloud, attack, defense, unsupervised learning, semantic segmentation, 2D/3D.**

I. Introduction

Nowadays.

II. Related Work

A. Unsupervised Domain Adaptation

B. Point Cloud Segmentation

C. Point Cloud Attack

III. Methodology

In this section, we present an overview of our Point Cloud Semantic Segmentation (PCSS) model's overall framework. Subsequently, we introduce a formal defense strategy built upon the foundational white-box attack framework.

A. Problem Definition and Notations

Let $\mathcal{S} = \{\mathcal{X}_S^{2D}, (\mathcal{X}_S^{3D}, \mathcal{Y}_S^{3D})\}$ be the source dataset that is composed of labeled point clouds and unlabeled pictures, where $\mathcal{X}_S^{3D}$ is a point cloud and $\mathcal{Y}_S^{3D}$ is its point-level labels, while $\mathcal{X}_S^{2D}$ is a picture without labels. Let $\mathcal{T} = \left\{\mathcal{X}_T^{2D}, \mathcal{X}_T^{3D}\right\}$ be the target dataset that is composed of unlabeled point clouds and unlabeled pictures, where $\mathcal{X}_T^{3D}$ is target point clouds, while $\mathcal{X}_T^{2D}$ is a target picture.

Regarding the attack strategy, 3D data $(\mathcal{X}_S^{3D}, \mathcal{Y}_S^{3D})$ is initially employed in a supervised manner within the attack framework, undergoing several iterations of training to generate adversarial samples. Subsequently, these adversarial samples are superimposed onto the 3D data of the target domain. This process disrupts the semantic segmentation framework's ability to accurately perform semantic segmentation, thereby achieving the attack.

Concerning the defense strategy, within the newly introduced 2D head, we utilize 2D data from the source domain as input to a 2D network. Feature extraction is performed, and an adversarial strategy is employed to deceive the discriminator, compelling the image features from both domains to possess similar distributions. By introducing 2D data, the model's robustness is enhanced, reducing the sensitivity of the 3D component to attacks and thereby achieving defense.

To this end, We first consider attacking the source domain point cloud data in the UDA framework of Cosmix [paper]. In the context where the source domain 3D point cloud data, point cloud category labels, and source domain model parameters are all ascertainable, adversarial training is conducted to generate adversarial samples. Subsequently, these adversarial samples are superimposed onto the target domain to execute the attack. Following this, an additional 2D semantic segmentation head is integrated into the original framework, analogous to the installation of a camera on a vehicle in real-world scenarios. Through the fusion and processing of data from multiple sensors, effective defense against such attacks can be achieved.

B. Attack Formulation

Below we formalize the attack goals as generalized used in adversarial attack.[paper]. The segmentation model is used to mapping the input point cloud $X = \{x_i | i = 1 \ldots N, x_i \in \mathcal{X}_S^{3D}\}$ to the labels of all points $Y = \{y_i | i = 1 \ldots N, y_i \in \mathcal{Y}_S^{3D}\}$. $\mathcal{X}$ is the real-time points sampled by the LiDAR system and sent to the semantic segmentation system, while $\mathcal{Y}$ is a set of feasible labels using in classification. We have established methodologies based on norm-bounded principles for both targeted and untargeted attacks, highlighting potential limitations in the attack scenarios. It is worth noting that encountering limitations in attacks is a commonplace occurrence in real-life situations.

C. Thread Model

1. Adversary's goals. We examine the LiDAR spoofing attack, assuming the attacker has white-box access to both the machine learning model and the perception system. The attacker's objective is to manipulate the semantic segmentation results obtained from PCSS models deployed in autonomous proxies such as vehicles and delivery robots, occurring in both indoor or outdoor scenarios.

The attacker is inclined to pre-train and employ adversarial samples for executing the attack, with limitations on the attack's effectiveness in cases where the attacker aims for imperceptibility. We find this threat model plausible, as attackers could gain access to the perception system and source data through additional engineering efforts, such as reverse engineering the software via the Internet or hardware hacking.

In this paper, we delve into two types of attacks: Targeted Attacks and Untargeted Attacks, both of which have proven to be effective in prior studies.

2. Targeted Attack. In this scenario, the attacker has access to both the target data and the semantic segmentation results for the target domain. This exposes information about the target objects, including their specific informations like relative location. Consequently, the attacker can selectively choose a specific subset of attack points denoted as $X_{\mathcal{T}} = \{x_i | i \in \mathcal{T}, x_i \in \mathcal{X}\}$. Here, $\mathcal{T}$ represents the indices of the items that the attacker intends to manipulate the predicted labels for, denoted as $Y_{\mathcal{T}} = \{y_i | i \in \mathcal{T}, y_i \in \mathcal{Y}\}$. The indices in $\mathcal{T}$ are randomly selected before the adversarial training process, assuming the attacker does not possess global information about the entire point cloud, thereby generating adversarial samples.

For a point $x_i = \{p_i\}$, we assume that the attacker conducts the model attack by perturbing the coordinates of the point, subject to necessary constraints. This perturbation involves the generalized investigation of point cloud coordinate perturbations. The perturbation values on the original points are represented as $R = \{r_i | i \in \mathcal{T}\}$, and the new point cloud is defined as $X' = \{x_i | i \notin \mathcal{T}, x_i \in \mathcal{X}\} + \{x_i + i_r | i \in \mathcal{T}, x_i \in \mathcal{X}\}$, where $r_i = r_{p_i}$ represents the coordinate perturbation on the 3D coordinates.

In this case, the attacker tries to minimize the difference between the predicted labels on $X_{\mathcal{T}}$ and the targeted labels Y_T, while the perturbation is bounded by ϵ. The attack goal can be formalized as:

$$\underset{R}{\operatorname{argmin}} \mathcal{L}_T(X', Y_{\mathcal{T}}), \text{s.t.} \mathcal{D}(R) \leq \epsilon \tag{1}$$

where $D(\cdot)$ is the distance function measuring the magnitude of the perturbation R and $\mathcal{L}_T(\cdot)$ is the adversarial loss that measures the effectiveness of the attack.

3. Untargeted Attack. In this Scenario, the attack target does not have a specific target $Y_{\mathcal{T}}$, but affect the predict labels $F_\theta(X_{\mathcal{T}})$ of target attack points to be different from the ground-truth labels $Y_{\mathcal{T}}$ of all the target points in $X_{\mathcal{T}}$, we follow the settings of [paper] and noted the labels as $Y_{G\mathcal{T}}$. While in the untargeted attack, we adjust the the loss Equation as:

$$\underset{R}{\operatorname{argmax}} \mathcal{L}_{NT}(X', Y_{G\mathcal{T}}), \text{s.t.} \mathcal{D}(R) \leq \epsilon \tag{2}$$

where $\mathcal{L}_{NT}(\cdot)$ is the adversarial loss that relate with $Y_{G\mathcal{T}}$, For setting a differentiable Loss function with the constraint, we reformulate the Equation 2 to enable gradient-based optimization by introducing $\mathcal{L}_{NT}(X', Y_{G\mathcal{T}})$ to replace the constraint. A smooth regularization $S(X')$ is set to make the perturbation point more imperceptible.

D. The Defense Strategy

IV. Experiments

V. Conclusions

References

[1]
[2]
[3]
[4]
[5]
[6]
[7]