Appendix

A Cluster-based Consistency Measurement

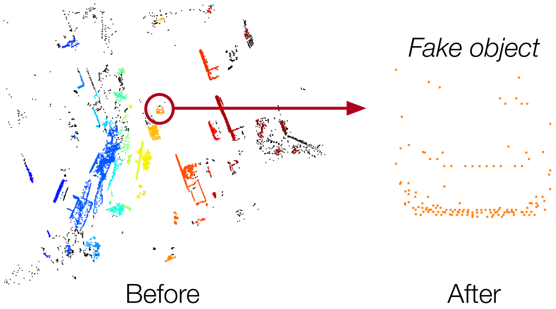


Figure 1: Qualitative Results from Spatial Clustering. Using predetermined thresholds, we identified 53 distinct clusters, each represented by a unique color, with black indicating outliers.

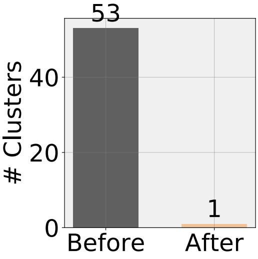


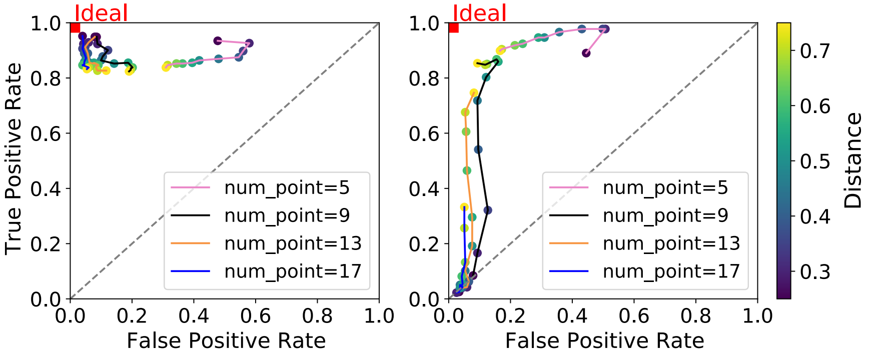
Figure 2: Cluster Count Changes. Removing clusters that contain synthesis points leaves a single cluster, indicating the attacker's counterfeit object.

We propose a cluster-based consistency measurement grounded in the fundamental concept that an object comprises more than one cluster and that point clusters attributed to a fake object lack consistency with clusters in historical frames. In this section, we show the visualization results (Figure 1) and cluster count changes (Figure 2) when applied to a poisoned frame where an attacker injects a point set crafted to mimic a car. `ADoPT` can identify fake objects in 97.2% of cases as shown in D.CAR in Table 1 in Sec. 6.1. Conversely, when our method is applied to benign frames, the absence of remaining clusters signifies the maintenance of object consistency across consecutive frames.

B Optimal Thresholds for Anomaly Detection

In determining the presence of anomalies, we apply DBSCAN, a prevalent spatial clustering technique, to all the points derived from merging the warped synthesis and the incoming frame. DBSCAN, by its nature, is a region-growing method that begins with an initial point and decides whether each subsequent point should be included in existing clusters or a new cluster. This decision-making process is significantly influenced by two threshold values: the minimum point requirement to form a cluster (count threshold) and the maximum distance between two points allowing them to belong in the same cluster (distance threshold). Therefore, determining an optimal threshold set that considers both the false positive rate and the true positive rate is essential for detecting various attacks.

However, determining the ideal threshold set proves to be challenging due to the varied levels of sparsity exhibited by the attacks, as depicted in Figure 3. The figure underscores the diverging trends in distance thresholds between the two types of attacks at a specific count threshold. For dense point injection, shorter distance thresholds are preferred, while sparse point injections benefit from longer distance thresholds.



(a) Dense Point Injection (b) Sparse Point Injection

Figure 3: Determining Optimal Clustering Thresholds Based on False Positive (FPR) and True Positive Rates (TPR). Each colored line signifies a unique count threshold, while each colored dot represents a distinct distance threshold. Every red square marks the ideal scenario where the TPR is 1 and the FPR is 0. In determining the optimal threshold, proximity to the coordinate of the ideal case serves as a guiding factor.

We determine the optimal thresholds by considering the distance from the coordinate of the ideal case where TPR is 1 and FPR is 0. For dense point attacks, the optimal threshold is achieved with a count threshold of 17 and a distance threshold of 0.25. This threshold leads to an FPR of 4.5% and a TPR of 95.2% when a fake pedestrian is injected. In the case of sparse point injection attacks, the optimal threshold is obtained with a count threshold of 9 and a distance threshold of 0.75, resulting in an FPR of 9.3% and a TPR of 85.4%.

C Failure Case Example

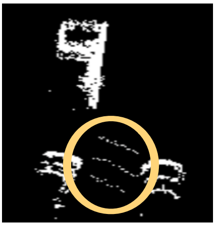


Figure 4: Representative Failure Case. The car marked with a yellow circle is a fake object created through a sparse injection attack. The perception module recognizes this fake car as part of the genuine vehicle on the right.

Using spatial clustering for attack detection predominantly fails when spoofed points are near benign road objects (see Figure 4). While this leads to a false negative, it does not markedly influence driving decisions or result in the failure to trigger true alarms, considering the imperative to avoid collisions with the benign object that remains in place.