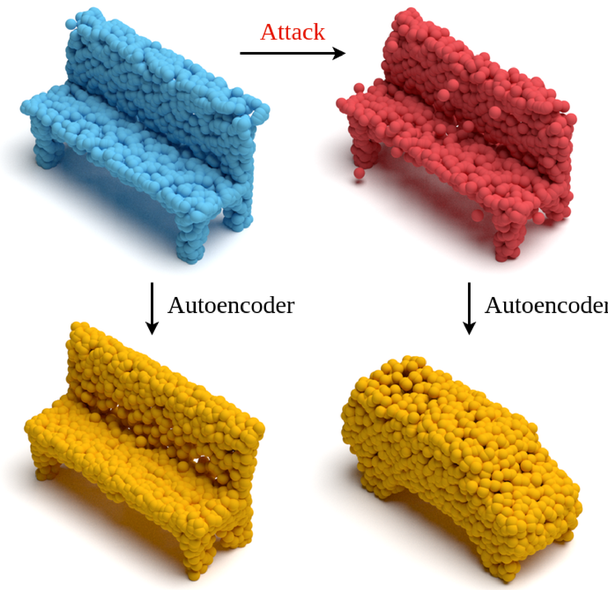
Geometric Adversarial Attacks and Defenses on 3D Point Clouds

Itai Lang, Uriel Kotlicki, Shai Avidan

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

3DV 2021
International Conference on 3D Vision
Online, December 1-3, 2021

Concept



Make a small **perturbation** to a clean source **point cloud** to change the reconstructed **geometry** by an autoencoder model

Contributions

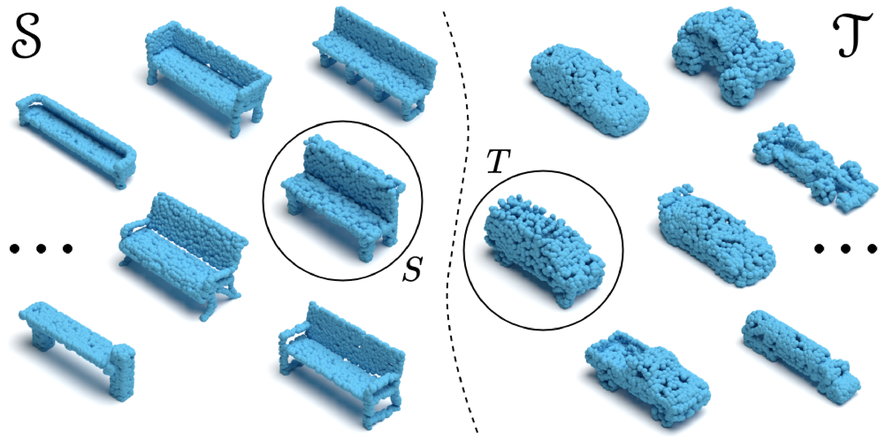
- The first to explore the problem of adversarial attacks at a geometric level
- Change the reconstructed point cloud to a shape instance from a different target class
- In sharp contrast to existing attacks on point cloud classifiers, which aim to alter the perceived semantic meaning of the object
- Extensive evaluation demonstrates the effectiveness of the proposed attacks and their robustness to counter defenses

Implementation

Our code is available!

Method

Target Selection

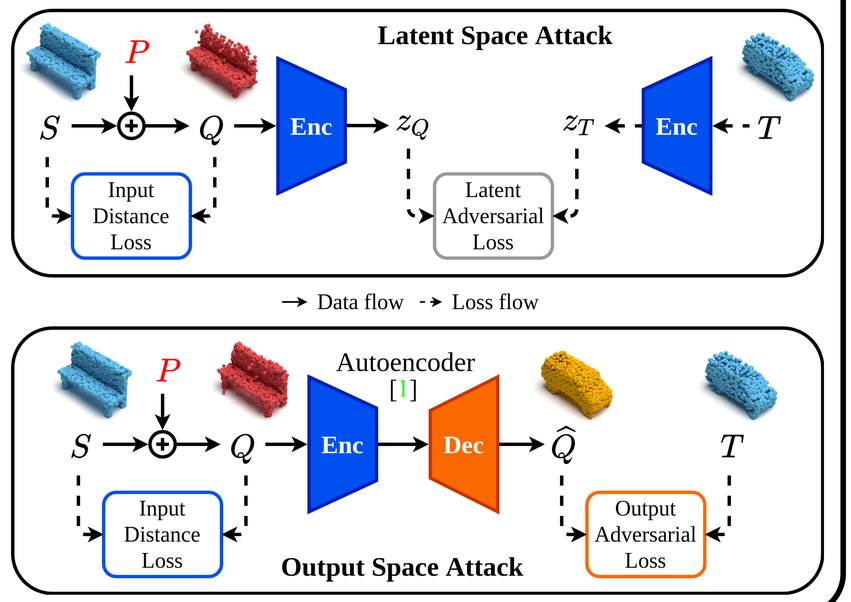


Latent space attack: $P^* = \underset{P}{\operatorname{argmin}}\ \mathcal{L}_{latent}(z_Q, z_T) + \lambda \mathcal{L}_{distance}(Q, S)$

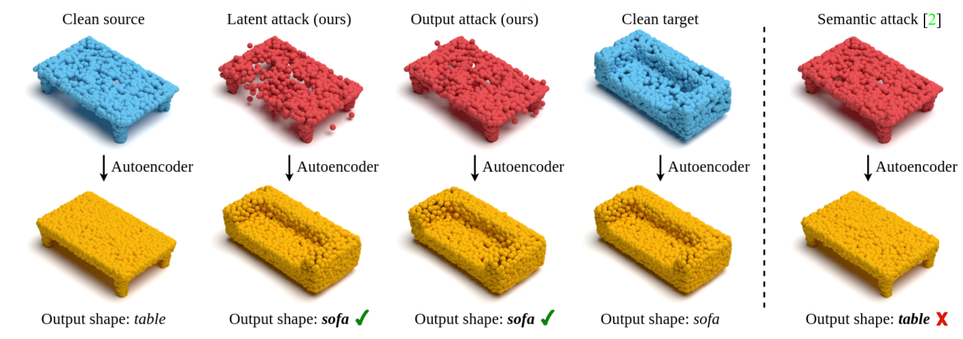
Output space attack: $P^* = \underset{P}{\operatorname{argmin}}\ \mathcal{L}_{output}(\widehat{Q}, T) + \lambda \mathcal{L}_{distance}(Q, S)$

Losses: $\mathcal{L}_{latent} = ||z_Q - z_T||_2 \quad \mathcal{L}_{output} = CD(\widehat{Q}, T) \quad \mathcal{L}_{distance} = CD(Q, S)$

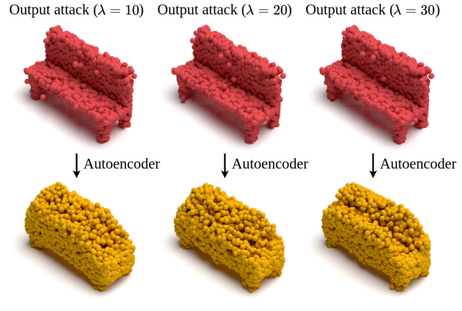
The Proposed Attacks



Attack



Both our attacks alter the reconstructed geometry to the desired target shape, while the output attack results in a smaller perturbation. In contrast, the semantic attack is ineffective against the autoencoder.



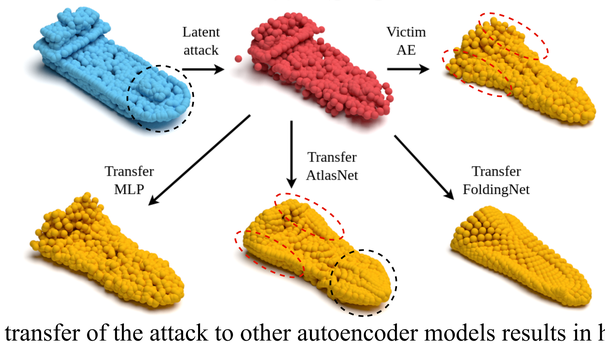
The attacks offer a trade-off between the source distortion and target reconstruction error. Even with the decreased perturbation, the geometry is changed.

Semantic Interpretation

Input Type	Hit Target	Avoid Source
Clean target	90.9%	98.1%
Semantic attack [2]	1.0%	9.6%
Latent attack (ours)	59.0%	86.3%
Output attack (ours)	**76.0%**	**94.7%**

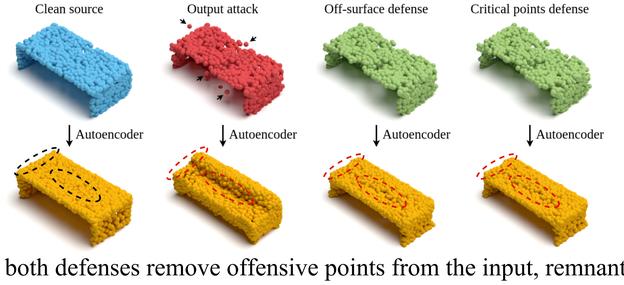
As a side effect, our geometric attacks also induce a semantic attack on the decoder side. That is, the reconstructions of our adversarial examples mislead a classifier to the desired target label or avoid the class of the source.

Transfer



The transfer of the attack to other autoencoder models results in hybrids of the source and target shapes, indicating some level of transferability.

Defense



While both defenses remove offensive points from the input, remnant geometric features of the attack's target shape are still present at the output.

[1] Achlioptas *et al.*, Learning Representaions and generative models for 3D Point Clouds, *ICML 2017*.
[2] Xiang *et al.*, Generating 3D Adversarial Point Clouds, *CVPR 2019*.