# Transferable 3D Adversarial Shape Completion using Diffusion Models

DAI Xuelong, XIAO Bin

Presenter: DAI Xuelong

August 21, 2024

# Table of Contents

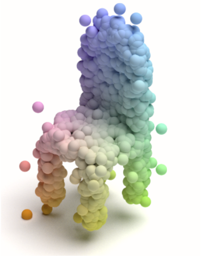
# Overview

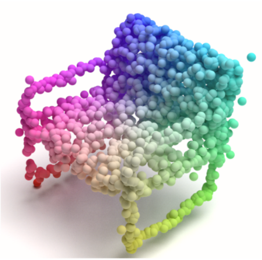
- We generate adversarial examples through shape completion using diffusion models against black-box point cloud models.
- We propose a variety of strategies to enhance the transferability of the proposed attacks without compromising the quality of generation.
- We conduct a comprehensive evaluation against existing state-of-the-art black-box 3D deep-learning models and achieve state-of-the-art performance against both black-box models and defenses.

# Table of Contents

# Motivation

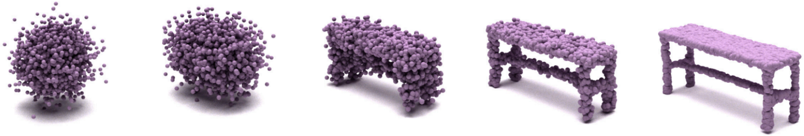
- 3D deep-learning models exhibit vulnerability to adversarial attacks, even when using 2D adversarial approaches.
- However, the perturbations applied to 3D point clouds that shift coordinates lead to noticeable changes in the original shape of 3D objects.





# Motivation

- Recent advancements in diffusion models applied to 3D point clouds have showcased remarkable performance in terms of both generation quality and diversity.
- These generative models can be utilized to generate high-quality unrestricted 3D adversarial examples.



# Motivation

- Moreover, existing 3D adversarial attacks face challenges in being effective against recently proposed state-of-the-art 3D deep-learning models, resulting in a huge gap in the development between adversarial attacks and benign models.

# Table of Contents

# Contribution

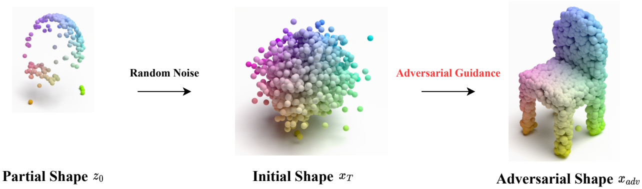
- We offer a novel perspective on the creation of imperceptible adversarial examples by using shape completion with diffusion models. The proposed attack introduces diffusion models to the topic of 3D adversarial robustness.
- We propose three effective techniques to improve the attack performance:
  - Employing model uncertainty for improved inference of predictions,
  - Ensemble adversarial guidance to boost attack performance against unseen models
  - Generation quality augmentation to identify critical points and maintain the quality of generation.
- We achieve effective attacks against recently proposed state-of-the-art 3D point cloud classifiers.

# Table of Contents

# Method: Adversarial Shape Completion

- We utilize any pre-trained 3D shape completion diffusion model $\epsilon_\theta$ to gradually generate the completed adversarial point cloud $x_0 = (z_0, \tilde{x}_0)$ through the reverse generative process $p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0)$, $t = T, \ldots, 1$.

$$p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0) := \mathcal{N}(\tilde{x}_{t-1} : \mu_\theta(x_t, z_0, t), \beta_t \mathbf{I}) - a\beta_t \nabla_{x_t} \mathcal{L}(f(x_t), y) \quad (1)$$



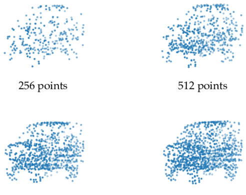
# Method: Adversarial Shape Completion

- In order to improve the effectiveness of the proposed attack on a black-box target model, we have outlined several effective strategies to enhance the transferability of the generated 3D point clouds.

# Method: Employing Model Uncertainty

- The removal of some points does not alter the classification outcome of the original point cloud.
- We are able to straightforwardly adopt model uncertainty to 3D deep-learning models with the *MC dropout*-like approach over the input.

$$\nabla_{x_t}\mathcal{L}_{\mathrm{MU}}(f(x_t), y)) = \frac{1}{M}\sum_{s=1}^{M}\nabla_{x_s}\mathcal{L}(f(x_s), y) \tag{2}$$



256 points 512 points 768 points 1024 points

# Method: Ensemble Adversarial Guidance

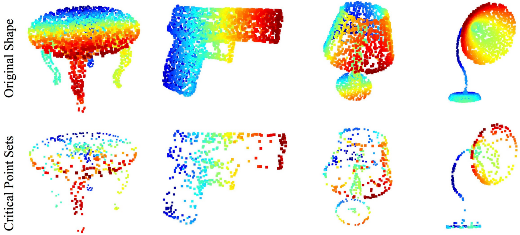
- The ensemble attack is an effective way to enhance the attack transferability.
- We ensemble the logits of selected substitute models according to the generative process.

$$\mathcal{L}(f_{ens}(x_t), y) = -\log(\text{softmax} \sum_{n=1}^{n_{\text{ens}}} w_n p_{f_n}(y|x_t)) \tag{3}$$

# Method: Generation Quality Augmentation

- Identifying critical points within the point cloud could achieve strong adversarial attacks.
- It is advisable to control perturbations by constraining the $\ell_0$ distance between the adversarial and benign point clouds.

$$\text{score}_x = \sum_3 \frac{\partial \mathcal{L}(f(x_t), y)}{\partial x} \tag{4}$$



# Table of Contents

# Evaluation

- Dataset: ShapeNetCore with 55 categories and 42003 data.
- Benign Diffusion Model: PVD.
- Comparisons:
  - White-box attacks: PGD, KNN, GeoA3, and SI-Adv
  - Black-box attacks: AdvPC, and PF-Attack
- Implementation: 3DAdvDiff to denote the white-box version and $\text{3DAdvDiff}_{\text{ens}}$ for boosting transferability version.

# Evaluation

| Method | PointNet | PointNet++ | DGCNN | PointConv | CurveNet | PCT | PRC | GDANet | Average |
|---|---|---|---|---|---|---|---|---|---|
| PGD | 99.9 | 2.1 | 0.7 | 0.8 | 0.5 | 0.4 | 0.7 | 1.6 | 0.9 |
| KNN | 99.9 | 2.2 | 0.7 | 0.7 | 0.5 | 0.6 | 1.1 | 1.6 | 1.1 |
| GeoA3 | 99.8 | 2.0 | 1.5 | 1.4 | 0.9 | 0.6 | 0.9 | 1.1 | 1.2 |
| SI-Adv | 92.5 | 2.0 | 1.7 | 1.5 | 1.2 | 1.0 | 1.3 | 1.0 | 1.4 |
| AdvPC | 89.6 | 0.4 | 0.2 | 0.5 | 0.4 | 0.6 | 0.7 | 0.5 | 0.5 |
| PF-Attack | 99.6 | 24.2 | 6.7 | 5.1 | 3.8 | 1.2 | 2.4 | 1.9 | 6.2 |
| 3DAdvDiff | 99.9 | 73.2 | 12.6 | 55.3 | 40.5 | 32.6 | 25.9 | 16.0 | 36.6 |
| 3DAdvDiff$_{ens}$ | 99.9 | **97.0** | ***99.9*** | **94.5** | **93.5** | **80.5** | ***99.9*** | **85.2** | **90.1** |

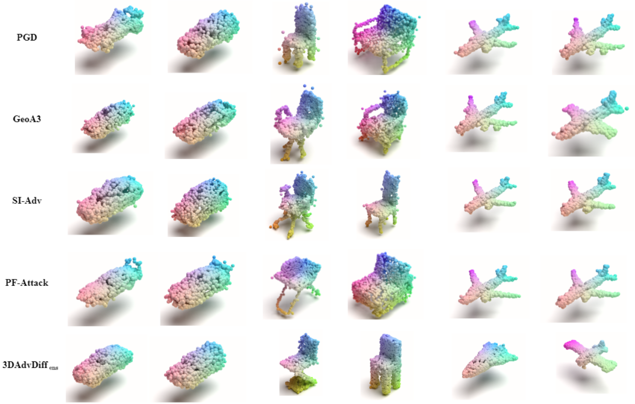
- The adversarial examples from state-of-the-art attacks merely transfer to different models, particularly those recently developed 3D models.
- Our proposed attack significantly outperforms exisiting attack methods.

# Evaluation

| Method | ASR | SRS | SOR | DUP-Net | IF-Defense | HybridTraining |
|---|---|---|---|---|---|---|
| PGD | 99.9 | 5.9 | 1.0 | 0.7 | 13.8 | 1.9 |
| KNN | 99.9 | 4.0 | 0.9 | 0.4 | 13.0 | 1.3 |
| GeoA3 | 99.8 | 4.9 | 1.6 | 0.8 | 13.6 | 2.2 |
| SI-Adv | 92.5 | 10.8 | 0.9 | 0.9 | 14.9 | 2.0 |
| AdvPC | 89.6 | 4.1 | 1.5 | 0.7 | 13.2 | 1.9 |
| PF-Attack | 99.6 | 8.5 | 3.6 | 2.8 | 13.9 | 2.0 |
| 3DAdvDiff | 99.9 | 82.2 | 9.9 | 9.6 | **30.0** | 9.4 |
| 3DAdvDiff$_{ens}$ | 99.9 | **85.9** | **49.1** | **36.9** | 22.5 | **96.1** |

- Existing defenses fail to defend against our attacks.
- Due to its utilization of model uncertainty, 3DAdvDiff is particularly effective against random sampling.
- The proposed critical point selection is effective against outlier removal defenses.

# Evaluation




- The visual quality of proposed attacks is comparable to existing attacks.

# Table of Contents

## Conclusion

- We introduce the first-ever method designed to execute a black-box adversarial attack on recently developed 3D point cloud classification models.
- We propose several strategies to effectively enhance the transferability of the proposed attack,
- Comprehensive experiments on the robust dataset validate the effectiveness of our proposed attacks.

# Q&A

Thank you for listening!

Your feedback will be highly appreciated!