Supplementary Materials for Transferable 3D Adversarial Shape Completion using Diffusion Models

Xuelong Dai[1] and Bin Xiao[1]

The Hong Kong Polytechnic University
xuelong.dai@connect.polyu.hk, b.xiao@polyu.edu.hk

1 ModelNet40

We report similar long-tail problems in the ModelNet40 [3] dataset as in Figure 1. To further validate the performance of the proposed 3DAdvDiff, we perform experiments on the ModelNet40 dataset. We select the top 8 classes to train the PVD model for shape completion with enough training data. The results are shown in Table 1 and 2. The proposed 3DAdvDiff$_{\text{ens}}$ outperforms existing attack methods remarkably on both black-box and against defenses.

Much like the ShapeNet dataset, black-box adversarial attacks typically perform poorly on categories within the ModelNet40 dataset that have a larger volume of training data. However, the test set of ModelNet40 is not uniformly selected. Instead of selecting a fixed proportion from the training data, ModelNet40 chooses 100 point clouds from all the top categories. As a result, the black-box Attack Success Rate (ASR) on ModelNet40 is relatively higher than that on the ShapeNet dataset. However, our proposed 3DAdvDiff$_{\text{ens}}$ still performs remarkably better than the previous attack methods.

2 Acceleration

The original PVD model adopts the DDPM [1] sampling for generating point clouds, which use 1000 sampling steps. An effective acceleration method to improve the sampling of DDPM is to use DDIM [2] sampling. We implement DDIM sampling for the PVD model with only 200 sampling steps. The results are shown in Table 3. We significantly improve the sampling speed without largely decreasing the generation quality. Improving time efficiency is a hot topic in the community, with many acceleration methods being introduced. Therefore, we believe the time efficiency of diffusion model based adversarial attacks can be further enhanced in the future.

3 Visual Results

We further give the visual results of our generated 3D adversarial point clouds in Figure 2 and 3.

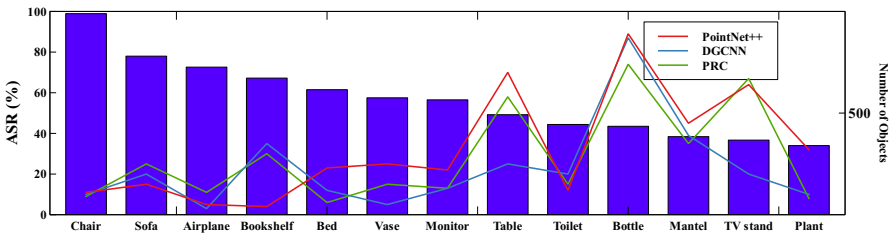

Fig. 1: The black-box ASR on the ModelNet40 dataset. We use the top 13 classes from the ModelNet dataset to demonstrate the long-tailed dataset problem. We use PGD with $\ell_{\text{inf}} = 0.16$ on PointNet to evaluate the black-box attack success rate (ASR).

4 Multi-View Adversarial Shape Completion

The shape completion tasks performed by the PVD model generate 20 different views for a specified partial shape to generate 20 different point clouds, which enables us to locate the most vulnerable views for generating adversarial point clouds. A similar finding is also addressed by Zhao et al. [4]. Therefore, it further enhances the performance of the proposed attacks as 3D deep learning models are sensitive to the transformations of 3D point clouds.

5 Adversarial Shape Generation

Diffusion models also have the ability to directly generate complete 3D point clouds without the need for a given partial shape. We further evaluated the performance of our proposed 3DAdvDiff in the context of adversarial shape generation, which we refer to as 3DAdvDiff-Gen. As demonstrated in Table 4 and 5, the 3DAdvDiff model, when used for adversarial shape generation, outperforms shape completion in black-box attacks, showing an average increase of 7.8% in Attack Success Rate (ASR). However, since shape generation does not inherently support multi-view generation during its original training, the white-box ASR is somewhat compromised without identifying the vulnerable transformations. Despite this, both time efficiency and training efficiency are enhanced. It should be noted, however, that the quality of adversarial shape generation is somewhat worse compared to shape completion. This could potentially be due to the absence of guidance on partial shape. We leave a better design of the shape generation in future works.

Table 1: The attack success rate (ASR %) of transfer attack on the ModelNet40 dataset. The adversarial examples of existing attack methods are generated from the PointNet model. The Average ASR is calculated among the seven black-box models ($3DAdvDiff_{ens}$ is calculated among the five black-box models).

Dataset	Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
Chair	PGD	99.9	11.6	9.2	27.5	4.5	15.2	7.1	9.0	12.0
	KNN	99.9	11.2	10.5	14.2	4.5	8.9	6.7	9.2	9.3
	GeoA3	99.9	7.1	4.6	8.1	3.5	5.2	6.5	4.9	5.7
	SI-Adv	99.9	65.4	37.4	30.4	20.2	13.8	22.8	19.4	29.9
	AdvPC	99.9	5.1	2.6	17.2	2.4	6.2	4.2	6.8	6.4
	PF-Attack	96.8	17.5	19.7	42.4	15.2	10.1	8.9	16.0	18.5
	3DAdvDiff	99.9	85.4	60.8	70.6	32.1	40.8	50.6	38.9	54.2
	$3DAdvDiff_{ens}$	99.9	**99.8**	***99.9***	**99.5**	**98.7**	**90.4**	***99.9***	**94.8**	**96.6**
Dataset	Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
All	PGD	99.9	22.8	18.8	54.5	18.3	16.1	15.4	23.4	24.1
	KNN	99.9	29.6	23.4	27.5	22.3	25.9	26.5	24.5	25.7
	GeoA3	99.8	15.3	9.5	15.6	10.3	10.6	12.4	10.8	12.1
	SI-Adv	99.9	60.4	30.1	58.4	34.8	36.8	24.5	36.8	40.2
	AdvPC	99.9	12.4	9.3	23.4	9.5	10.1	8.3	10.8	12.0
	PF-Attack	99.1	51.4	31.3	67.4	35.4	35.4	23.1	41.2	40.7
	3DAdvDiff	99.9	90.1	65.8	85.4	52.8	63.9	51.8	67.4	68.2
	$3DAdvDiff_{ens}$	99.9	**99.8**	***99.9***	**99.9**	**98.6**	**91.2**	***99.9***	**96.3**	**97.2**

Table 2: The attack success rate (ASR %) of different adversarial attack methods against defenses. All attacks are evaluated under white-box settings against the PointNet model.

Method	ASR	SRS	SOR	DUP-Net	IF-Defense	HybridTraining
PGD	99.9	61.3	17.6	16.5	14.3	0.4
KNN	99.9	94.5	85.4	48.9	22.4	12.1
GeoA3	99.8	55.3	28.6	22.1	13.6	1.5
SI-Adv	92.5	75.1	22.1	20.3	18.6	19.1
AdvPC	99.9	84.8	21.4	19.8	20.6	0.4
PF-Attack	99.1	47.5	77.3	43.0	29.2	13.6
3DAdvDiff	99.9	95.6	90.5	88.3	**52.1**	31.5
$3DAdvDiff_{ens}$	**99.9**	**98.7**	**96.0**	**95.4**	43.7	**98.6**

6 Limitation

Given the unique characteristics of 3D point clouds, they necessitate a larger volume of training data compared to 2D images when training diffusion models. At present, all existing 3D diffusion models are trained using the large-scale classes in the ShapeNet dataset. This, however, restricts the generalizability of the proposed diffusion adversarial attacks to relatively smaller datasets. Nonetheless, we are optimistic that with the continued advancement of 3D diffusion models, a large-scale and balanced 3D dataset will become available in the future. Furthermore, while we have managed to enhance the sampling speed of our proposed 3DAdvDiff with DDIM sampling, the generation speed of the proposed attack still lags behind PGD-based attacks. However, with the rapid development of diffusion models, the time efficiency problem is addressed in many recent works.

Table 3: The attack performance with DDPM and DDIM sampler. All attacks are evaluated under white-box settings against the PointNet model on all selected classes of the ShapeNet dataset.

	ASR	Time(s)	CD
DDPM	**90.1**	60.8	**0.14**
DDIM	89.9	**13.5**	0.18

Table 4: The attack success rate (ASR %) of adversarial shape completion and shape generation. The adversarial examples of existing attack methods are generated from the PointNet model on the ShapeNet's Chair class. The Average ASR is calculated among the seven black-box models.

Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
3DAdvDiff	**99.9**	**60.6**	8.7	23.5	9.8	6.9	14.9	8.9	19.0
3DAdvDiff-Gen	90.1	54.2	**25.4**	**32.1**	**21.2**	**7.6**	**24.5**	**22.5**	**26.8**

Our future goal is to further boost the efficiency of 3DAdvDiff by incorporating acceleration techniques from diffusion models.

7 Ethics Concerns

The proposed 3DAdvDiff brings new challenges to 3D deep learning models. Adversaries may adopt our attacks to generate malicious point clouds to attack the 3D deep learning classification models. However, our proposed 3DAdvDiff can also be utilized for adversarial training to enhance the robustness of 3D deep learning models. The proposed attack can further encourage the development of 3D adversarial defenses. Therefore, our proposed 3DAdvDiff can achieve positive impacts on improving the 3D deep learning model robustness.

References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
2. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
3. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
4. Zhao, Y., Wu, Y., Chen, C., Lim, A.: On isometry robustness of deep 3d point cloud models under adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1201–1210 (2020)

Table 5: The generation quality on the ShapeNet dataset. The CD distance is multiplied by 10^{-2}.

Method	3DAdvDiff$_{ens}$	3DAdvDiff$_{ens}$-Gen
HD	**0.098**	0.80
CD	**0.14**	0.36
MSE	**1.18**	3.05

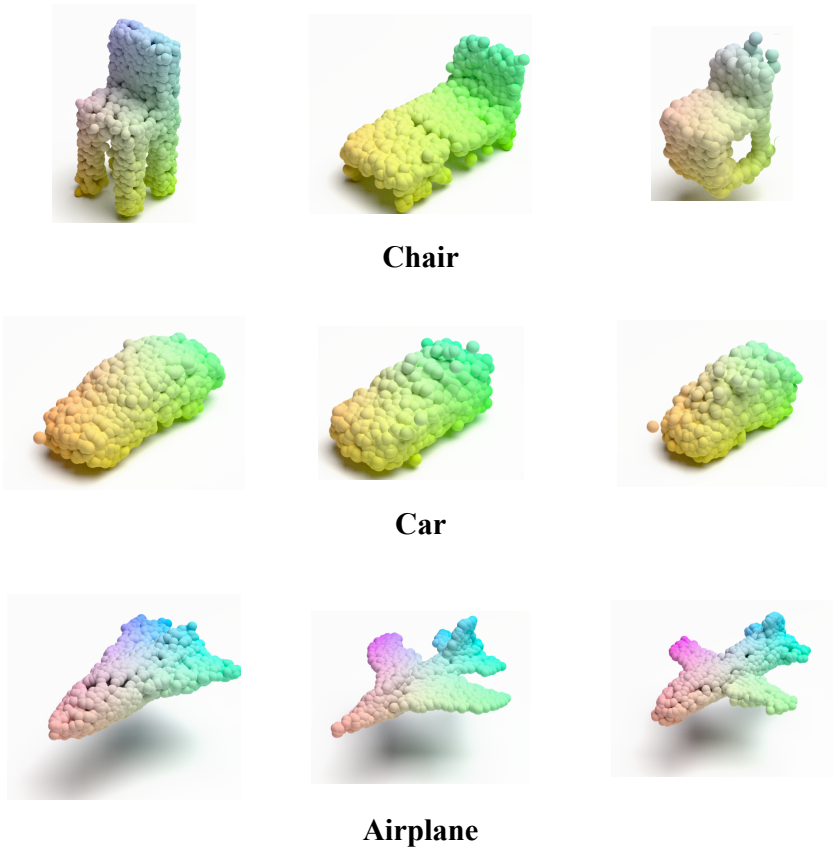

Fig. 2: The generated adversarial point clouds of 3DAdvDiff$_{ens}$.

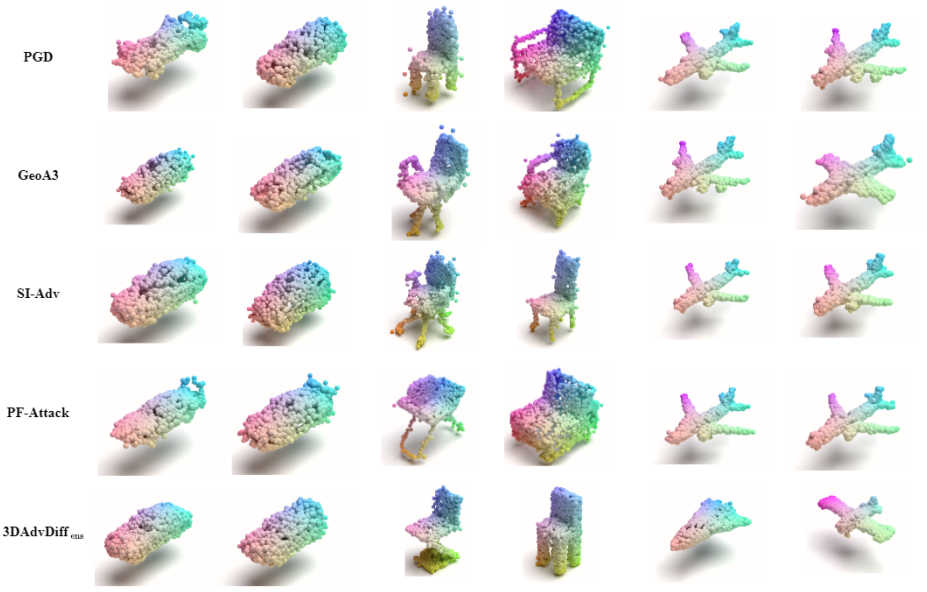

Fig. 3: The generated adversarial point clouds. The adversarial examples are randomly sampled from ShapeNet dataset.