Towards Transferable Targeted 3D Adversarial Attack in the Physical World

Supplementary Material

1. Physical Implementation Details

For the implementation of physical attacks, here we detail the specific process including the used equipment. The whole implementation process could be split into three stages: Preprocess—3D Print—Capture and Test.

Preprocess. Before 3D printing, there is still a need for some preprocessing of the generated adversarial 3D mesh. This is because we have parameterized the appearance information into the parameters of grid-based NeRF. Hence, it is necessary to convert this information back into the texture maps required for 3D printing. Specifically, to extract the appearance as texture images, we first unwrap the UV coordinates of $\mathcal{M}_{adv}$ using XAtlas [2]. Subsequently, we bake the surface's color into an image of the texture map corresponding to the UV coordinates, which could be used for the following 3D printing.

3D Print. Then, we print the 3D adversarial examples generated by TT3D in the form of textured meshes using well-established 3D printers. In our case, we utilize the J850™ Digital Anatomy™ 3D Printer[1], a classic 3D printer known for its versatility and precision. This printer can accurately reproduce both the textures and geometries of our adversarial samples, offering a wide range of material choices and color options. Here, we print a total of 20 objects, including 10 against the surrogate model ResNet-101 and 10 against the surrogate DenseNet-121.

Capture and Test. In the last stage, we aim to validate the performance of the 3D adversarial samples generated by TT3D across different viewpoints and backgrounds in the physical world. Thus, as mentioned in the manuscript, we place the 3D adversarial object on the given surface (with different backgrounds B-1, B-2, and B-3) and slowly circle them with a smartphone for about 360°, excluding the bottom part. which lasts approximately 20 seconds per object in each setting, capturing 10 frames per second, resulting in a total of 200 frames. Finally, we could calculate the attack success rate as the final test result, which is determined by the proportion of successful frames.

2. More Experiments

2.1. Choice for Epoch

To ensure convergence and prevent overfitting, selecting an appropriate epoch for TT3D is crucial. Thus, following the same settings as described in Section 4.1 of the manuscript, we have visualized the variation curve of the total loss value with respect to the epoch for the attack surrogate model ResNet-101 in Fig. 1. The figure reveals that when the epoch reaches the number of 250, the total loss value gradually stabilizes, signifying the achievement of convergence. To prevent overfitting, we have thus chosen 250 as the final epoch for optimization.

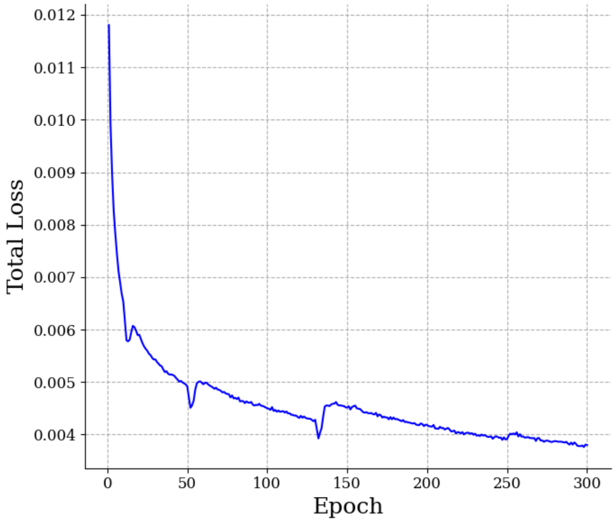

Figure 1. The total loss of TT3D with respect to the epoch against the surrogate model ResNet-101 with different training iterations.

2.2. Evaluation for Naturalness

To more objectively evaluate the naturalness of 3D adversarial samples generated by TT3D, three metrics for assessing image quality are employed here: Structural Similarity Index [1] (SSIM), Peak Signal-to-Noise Ratio (PSNR),

Methods \ Metrics		SSIM↑	PSNR↑	LPIPS↓
Initial		0.9821	39.00	0.0293
Mesh-based		0.8259	13.79	0.1912
TT3D(RN-101)	$\beta = 10^2$	0.8741	22.54	0.1447
	$\beta = 10^3$	0.9039	27.96	0.1153
	$\beta = 10^4$	0.9486	33.26	0.0874

Table 1. The results of quantitative naturalness metrics: **PSNR, SSIM, and LPIPS scores** for the initially reconstructed clean images, adversarial samples generated by the previous mesh-based method, and the corresponding samples produced by our TT3D method under $\beta = 10^2$, $\beta = 10^3$, and $\beta = 10^4$. Higher SSIM and PSNR values indicate better performance, while a lower LPIPS score is preferable.

[1] https://www.stratasys.com/en/3d-printers/printer-catalog/polyjet/j850-digital-anatomy/

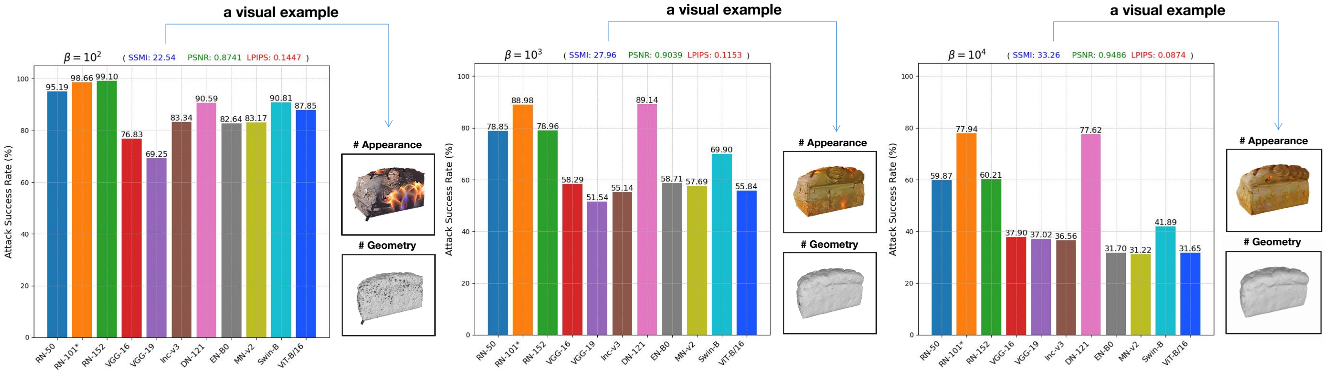

Figure 2. The **attack success rate(%)** of 3d adversarial examples generated by TT3D under $\beta = 10^2$, $\beta = 10^3$, and $\beta = 10^4$ against ResNet-50 (RN-50), ResNet-101 (RN-101), ResNet-152 (RN-152), VGG-16, VGG-19, Inception-v3 (Inc-v3), DenseNet-121 (DN-121), EfficientNet-B0 (EN-B0), MobileNet-v2 (MN-v2), Swin-B, and VIT-B/16. The 3d adversarial examples are learned against surrogate models ResNet-101. Each bar chart is accompanied by a visual example in the lower right position, showcasing the appearance and geometry of a 3d adversarial sample (attack 'bread' to 'stove') generated by TT3D under the corresponding β value.

Source Model	Background	Victim Model										
		RN-50	RN-101	RN-152	VGG-16	VGG-19	Inc-v3	DN-121	EN-B0	MN-v2	Swin-B	ViT-B/16
ResNet-101	B-1	45.65	81.30*	80.20	**31.25**	**37.90**	49.30	73.30	**33.45**	45.30	**71.75**	39.25
	B-2	42.35	76.45*	62.85	21.35	23.60	47.70	61.30	23.90	**59.65**	63.20	37.70
	B-3	**84.70**	**91.55***	**89.70**	29.60	32.80	**69.60**	**90.40**	32.35	36.60	68.60	**64.15**
DenseNet-121	B-1	39.25	34.70	36.35	29.70	26.55	32.15	79.30*	31.20	33.75	43.20	21.20
	B-2	**41.30**	**37.65**	**39.20**	35.50	41.70	34.90	83.25*	32.95	35.20	62.25	**31.85**
	B-3	39.45	35.95	37.60	**51.55**	**42.30**	**37.45**	**84.50***	**34.20**	**36.15**	**77.30**	29.80

Table 2. The **attack success rates(%)** of 3d printed adversarial objects with different backgrounds in the physical world against ResNet-50 (RN-50), ResNet-101 (RN-101), ResNet-152 (RN-152), VGG-16, VGG-19, Inception-v3 (Inc-v3), DenseNet-121 (DN-121), EfficientNet-B0 (EN-B0), MobileNet-v2 (MN-v2), Swin-B, and VIT-B/16. The 3d adversarial objects are learned against the surrogate model ResNet-101 and DenseNet-121 and produced with 3d printing techniques for physical implementation.

and Learned Perceptual Image Patch Similarity [3] (LPIPS). SSIM evaluates perceptual degradation by analyzing structural, luminance, and contrast changes. PSNR measures pixel-level accuracy between original and distorted images, with higher values indicating better fidelity. LPIPS, leveraging deep learning (vgg used here), evaluates perceptual similarity at the patch level, reflecting the human visual system's response to image variations. Specifically, we compare the above three quantitative metrics of the initially reconstructed clean images, adversarial samples generated by the previous mesh-based method, and the corresponding ones produced by our TT3D method under various β values. β is the weight of the regularization for naturalness in TT3D. Experimental results are listed in Tab. 1, revealing that our TT3D significantly outperforms the mesh-based method even under different β values and exhibits an acceptable decline compared to the clean images, thereby confirming the superior naturalness of TT3D.

2.3. Effects of β

As mentioned in the manuscript, β, an adjustable weight for the regularization $\mathcal{R}$ responsible for naturalness, has a significant impact on both naturalness and attack performance. To perform a more comprehensive analysis of β's effects, here we measure the attack success rate of TT3D (against the surrogate model ResNet-101) across various black-box classifiers with different β values of 10^2, 10^3, and 10^4. Additionally, we provide a visual example of the adversarial samples generated under these settings. Experimental results, as presented in Fig. 2, where we can see that: 1) **At $\boldsymbol{\beta = 10^2}$**, when β is relatively low, the success rate across different models is predominantly above 80%. However, this comes at the cost of a certain degree of naturalness, both in appearance and geometry. 2) **At $\boldsymbol{\beta = 10^4}$**, under a strong regularization weight, we still observe considerable success rates, with the lowest being above 30%. 3) **At $\boldsymbol{\beta = 10^3}$**, there is a relative balance between success rate and naturalness, with both metrics demonstrating objectively favorable outcomes. Consequently, this β value, i.e., $\beta = 10^3$ is chosen as the final implementation parameter for TT3D.

2.4. Transferability in the Physical World

Our TT3D could achieve a transferable targeted attack not only in the digital world but also in the physical world. The

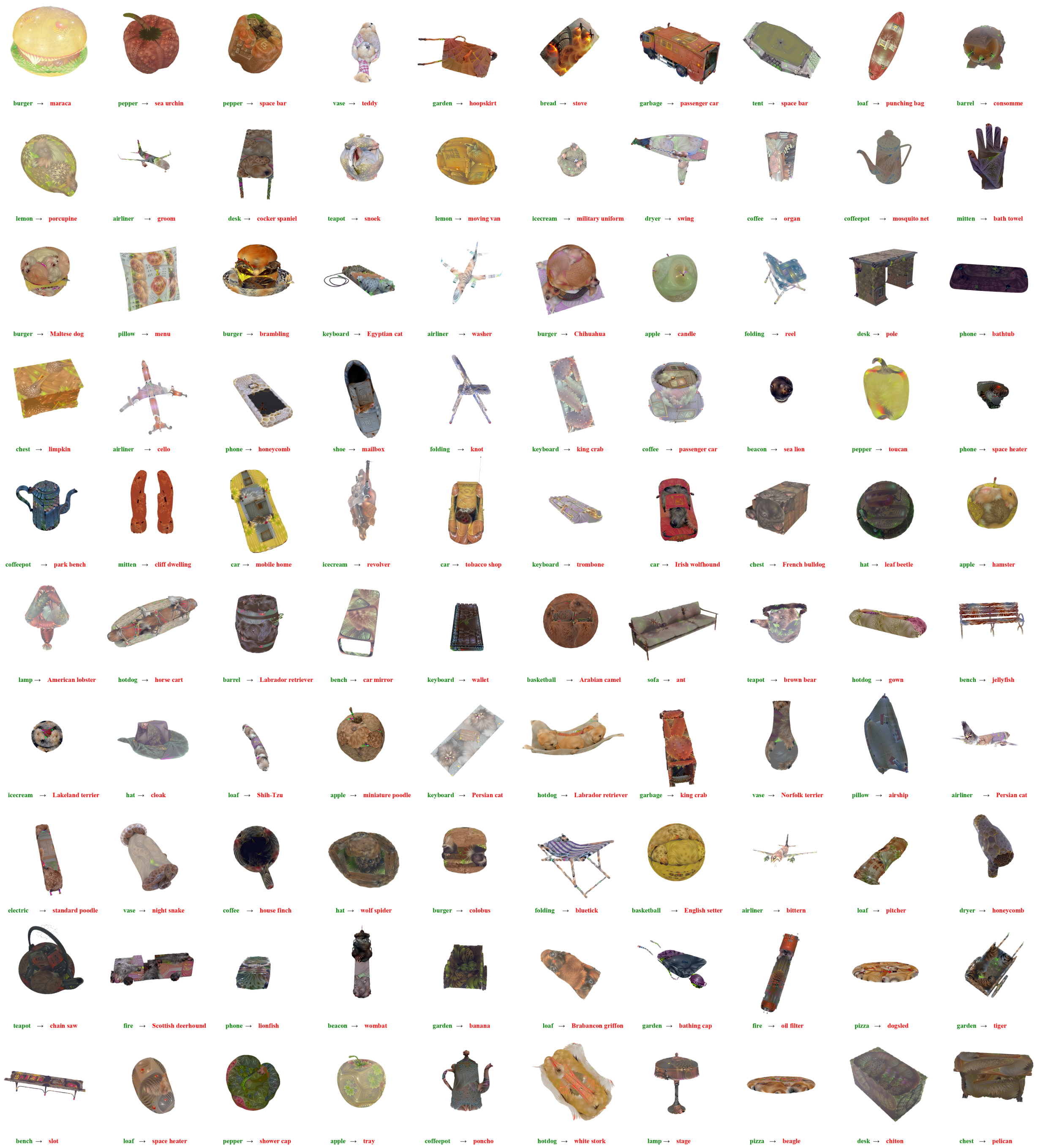

Figure 3. The total of 100 different 3D adversarial objects, generated by TT3D to attack the surrogate models ResNet-101. The above image of each object is captured under a single, random viewpoint, and in the experiments, we capture 100 different viewpoints for calculating the ASR. The green text under each object is the clean label, and the red text is the target label, which is also randomly chosen from the labels in Imagenet. The above images show the diversity of the object classes involved in the attack and the randomness of the target categories.

specific results in the digital world have been given in the manuscript. Here, to verify the performance in the physical world, we follow the implementation process in Sec. 1 and test our 3D-printed adversarial objects' attack success rate against various black-box classifiers in the physical world. As shown in Tab. 2, even when confronted with varying backgrounds, our 3D adversarial objects still maintain commendable attack success rates across a spectrum of black-

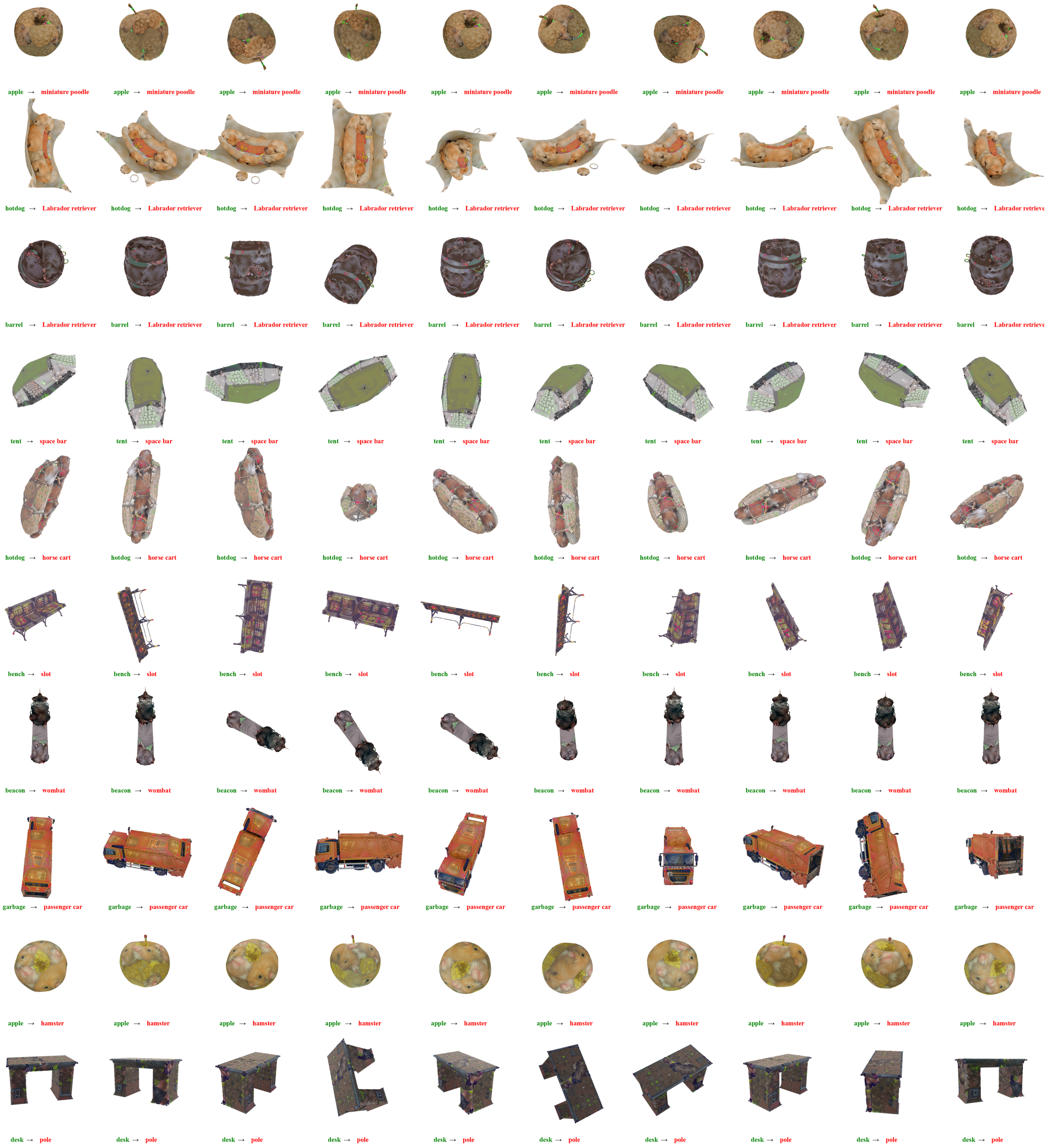

Figure 4. The total of 10 randomly selected 3D adversarial objects under 10 random viewpoints, along with their predicted outcomes. This validates the randomness and diversity of our viewpoint selection, as well as the robustness of our TT3D method across varying viewpoints.

box classifiers, which verifies the robustness of our TT3D.

3. More examples of TT3D

In this section, we present more 3D adversarial samples generated by TT3D: (1) Fig. 3 shows 100 different 3D adversarial objects, generated by attacking the surrogate model ResNet-101. These images, captured from a single random viewpoint, demonstrate the diversity of the object classes involved in the attack and the randomness of the target categories. (2) Fig. 4 consists of 10 randomly selected 3D objects from Fig. 3, each depicted from 10 different viewpoints (100 random viewpoints used in the experiments), to illustrate the effectiveness of the 3D adversarial samples under various viewpoints.

References

[1] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[2] Jonathan Young. xatlas, 2021. 1

[3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2