Allegory of the Cave: Breakdown of Illusions in Multimodal Perception with Neural Radiance Fields

Axel Päivänsalo*†, Ching-Chun Chang*‡, Hanrui Wang‡, Futa Waseda§, and Isao Echizen‡§
† Aalto University, Espoo, Finland E-mail: axel.paivansalo@gmail.com
‡ National Institute of Informatics, Tokyo, Japan E-mail: ccchang@nii.ac.jp
§ University of Tokyo, Tokyo, Japan

***Abstract*—Multimodal neural networks excel at linguistic understanding and visual recognition, yet they remain notoriously vulnerable to adversarial examples—inputs subtly perturbed with crafted noise that induce perceptual illusions and lead the model to misclassification. This vulnerability poses serious risks in safety-critical systems, where even imperceptible distortions may result in catastrophic decisions. In this study, we introduce a purification mechanism based on a perspective shift inspired by Plato's Allegory of the Cave. Rather than directly classifying potentially deceptive two-dimensional samples, we seek to uncover the underlying reality by reconstructing the physical scene in three dimensions. The purification process involves three-dimensional reconstruction with neural radiance fields and two-dimensional reprojection to align with the input space of multimodal perception. The resulting samples preserve physical consistency, with pixel-level perturbations removed. Our experiments demonstrate that the proposed mechanism significantly improves the robustness of multimodal AI against various types of adversarial attacks in comparison to benchmark methods. This work advances adversarial defense by moving beyond pixel-level denoising toward a deeper understanding of physically grounded coherence.**

I. Introduction

Modern deep learning models, though highly performant, are susceptible to adversarial examples—input samples perturbed in ways imperceptible to humans yet capable of misleading neural networks into making incorrect predictions. This fragility, first identified by [1], raises critical concerns for real-world applications such as autonomous driving and medical imaging, where erroneous decisions can lead to security breaches and catastrophic consequences.

A widely studied defense against adversarial attacks is adversarial training [2], which enhances robustness by augmenting the training data with adversarial examples. However, this approach is computationally intensive and prone to overfitting to specific attack patterns. An alternative line of defense applies input transformations and filtering at inference time to mitigate adversarial noise. Techniques such as JPEG compression, blurring, and bit-depth reduction [3] can be effective against simple perturbations, although they are often circumvented by adaptive attacks. More recently, generative purification strategies using deep generative models have shown considerable promise. For example, DiffPure [4] employs a diffusion model to denoise adversarial inputs through a forward-reverse generative process, achieving state-of-the-art performance against adversarial attacks. Nonetheless, these methods remain fundamentally constrained by their pixel-level perspective, making them vulnerable to attacks that exploit the limitations of this view. As a result, they may falter when faced with spatially concentrated perturbations, such as adversarial patches.

To address this limitation, we propose a fundamentally different perspective rooted in *Plato's Allegory of the Cave*. In essence, the allegory suggests that genuine understanding requires transcending the limitations of sensory perception through a perspective shift. Just as the prisoners in the cave mistake shadows for reality, deep networks operating solely on 2D pixels risk making decisions based on deceptive projections. We argue that robust perception requires emerging from this cave—that is, reconstructing and reasoning about the physical 3D structure that gives rise to the image. This philosophical shift motivates 3DPure, a novel adversarial defense that purifies images by reconstructing a 3D model of the scene and rendering a new, physically consistent 2D image for classification. By enforcing coherence with plausible 3D geometry, 3DPure filters out adversarial artifacts that have no structural basis in reality. This approach proves especially effective against spatially localized attacks such as adversarial patches, which are naturally disrupted during the reconstruction process.

We implement 3DPure using TripoSR, a fast feed-forward single-view 3D reconstructor, and evaluate the classification accuracy of the purified renderings with a pre-trained CLIP-based zero-shot vision-language model. Our contributions are threefold: 1) the introduction of 3DPure, the first defense to leverage 3D reconstruction from a single image to purify adversarial inputs by enforcing physical plausibility; 2) the development of a pipeline using TripoSR and CLIP, demonstrating compatibility with existing multimodal models; and 3) validation across multiple datasets and attack types, showing substantial gains in robustness, particularly against spatially localized patch-based attacks, compared to state-of-the-art defenses.

II. Related Work

The challenge posed by adversarial examples has spurred extensive research into developing robust defense mechanisms.

*Equal contributions.

979-8-3315-7206-8/25/$31.00

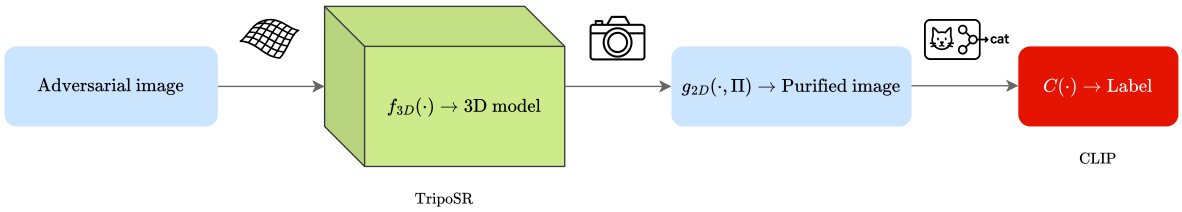

Fig. 1. High-level flow of 3DPure. An adversarial image (left) is first passed through a 3D reconstructor (TripoSR) that produces a 3D representation of the image. Re-rendering that model from a (possibly novel) viewpoint yields a purified 2D image whose pixels are now constrained by physical consistency. Finally, a conventional 2D classifier (CLIP) processes this image and recovers the correct label.

While many strategies focus on 2D image-space manipulations or refining model training protocols, an increasingly explored frontier involves leveraging 3D information and geometric understanding. These approaches aim to instill physical plausibility into the defense process, potentially offering more fundamental ways to mitigate adversarial perturbations.

A prominent line of research utilizes learned 3D-aware priors or exploits rendering consistency to filter adversarial inputs. A notable example is IF-Defense [5], which operates on 3D point clouds by learning to restore object surfaces through an implicit surface regularizer, effectively using a prior on natural 3D geometry to smooth out adversarial distortions. Extending this concept to the 2D domain, DISCO [6] adapts local implicit functions for image purification, treating an image as a continuous function and using a decoder to project adversarial inputs onto the manifold of natural images, thereby removing high-frequency noise. Beyond implicit surfaces, other methods explore 3D rendering techniques and multi-view consistency. For instance, inconsistencies in color for the same 3D points observed in stereo image pairs attacked independently can reveal adversarial manipulation [7]. Furthermore, emerging ideas consider technologies like Neural Radiance Fields (NeRFs). While primarily explored for generating multi-view consistent attacks [8], the concept of fitting a NeRF to a perturbed image and then re-rendering a "clean" version offers a promising 3D-aware filtering mechanism, as the rendered output is constrained by the learned 3D geometry and appearance.

Complementary to implicit and rendering-based methods, another strategy involves the use of explicit, parametric 3D models. These models, such as 3D Morphable Models for human faces [9] or generic CAD models for other object categories, provide strong priors on expected shape and texture. By fitting these models to an adversarial image and then re-rendering, the system can enforce known structural constraints. For example, fitting a 3DMM to a face image can help reconstruct a canonical view, effectively removing occlusions or adversarial patterns. This projects the image onto the space of valid appearances defined by the model, discarding artifacts that violate its geometric and textural degrees of freedom. Similarly, monocular depth estimation can act as a 3D prior; an estimated depth map, even from a perturbed image, can inform a re-rendering process that filters out perturbations inconsistent with the inferred 3D scene layout.

III. Methodology

Two mechanisms make 3DPure effective. First, the 3D reconstructor enforces a strong shape prior and implicitly regularizes appearance, since every rendered pixel must originate from a coherent 3D surface. Random pixel noise or adversarial patterns that do not correspond to a plausible surface on the object are discarded in this process. Second, we can optionally perturb the rendering viewpoint. This means an adversarial perturbation would have to simultaneously fool the classifier from multiple angles—a much harder task—thus undermining attacks like PGD or patches that rely on a fixed alignment. In practice, we found in Section IV that using the original view preserves accuracy best, so our main results use the original viewpoint unless stated otherwise.

Formally, 3DPure consists of a sequential pipeline of operations: an adversarial input image is mapped to a 3D representation, which is then rendered back to 2D for final classification (see Fig. 1). Let x_{adv} be an input image. We define a transformation $\Phi(x) = g_{2D}(f_{3D}(x))$, where f_{3D} is the 3D reconstruction function and g_{2D} is the rendering function. The purified image is $\tilde{x} = \Phi(x_{\text{adv}})$, which is then fed to a classifier C for prediction. We summarize this in Eq. (1):

$$\tilde{x} = g_{2D}\big(f_{3D}(x_{\text{adv}})\big), \quad \hat{y} = C(\tilde{x}) \,, \tag{1}$$

where $\hat{y}$ is the predicted label. The hope is that $\hat{y}$ equals the true label y even if x_{adv} was perturbed, because the purification Φ removes the adversarial perturbation.

A. 3D Reconstruction via TripoSR

We implement f_{3D} using the **TripoSR** model, introduced in 2024 by [10]. TripoSR is a state-of-the-art single-image 3D reconstruction network that predicts a neural radiance field (NeRF) representation of an object from one input image. In 3DPure, we use TripoSR without modification to convert a 2D image into a 3D mesh. As a pre-processing step, we apply background removal on the input image to focus reconstruction on the object. This improves the robustness of 3DPure by eliminating background clutter and background-based perturbations.

B. Rendering and Classification with CLIP

After obtaining the 3D object model, we render a 2D image $\tilde{x} = g_{2D}(M, \Pi)$, where M is the 3D mesh and Π is the

camera pose. The rendering produces a purified image that is expected to be free of adversarial noise (since M could not model the noise). Finally, we classify $\tilde{x}$ using the **CLIP** model introduced by [11]. CLIP (Contrastive Language-Image Pre-training) is a powerful vision-language model that encodes images and text into a shared embedding space. We choose CLIP as our classifier C for two reasons: (1) It is robust and flexible, having been trained on a broad dataset of image-text pairs. (2) It allows us to define the label space via text prompts, which is convenient for evaluating on multiple datasets. In our experiments, we use a pretrained CLIP (ViT-B/32) to predict labels by comparing the image embedding of $\tilde{x}$ with text embeddings of class names.

IV. Experiments

We extensively evaluate 3DPure on multiple datasets under various adversarial attacks, comparing its performance with baseline defenses. We describe our experimental setup below.

Datasets: We test on four datasets providing a mix of object types and complexities: (1) CUB-200-2011 (CUB) introduced by [12]—a fine-grained dataset of 200 bird species. CUB images typically depict clearly visible birds, fully contained within image borders without occlusion, making it highly suitable for evaluating fine-detail retention. We randomly sampled a consistent subset of 1,000 images from CUB for all evaluations and employed two prompting strategies: a simple prompt distinguishing broad categories ("bird", "dog", "cat", "car", "airplane", "zebra"), referred to as Birds, and a complex prompt involving all 200 bird species, testing fine-grained classification capabilities. (2) CO3D by [13]—a collection of Common Objects in 3D by Meta, containing multi-view images of 50 object categories. We use single-view images from CO3D to evaluate 3DPure's generalization on common objects. (3) Caltech101 by [14]—a classical dataset of 101 object categories, included to assess performance on a well-established benchmark. A total of approximately 1,000 images per dataset are used, all resized to a fixed resolution (224×224) for input to CLIP.

Attacks: We consider both ℓ_∞ norm-bounded adversarial noise and patch attacks. For ℓ_∞ attacks, we use: (1) FGSM (Fast Gradient Sign Method) by [1]: a single-step attack $x_{\text{adv}} = x + \epsilon\,\text{sign}(\nabla_x L(x, y))$ with $\epsilon = 8/255$, where $L(x, y)$ denotes the classifier's loss function given input x and true label y. This is a relatively weak attack but very fast. (2) PGD (Projected Gradient Descent) by [2]: a stronger iterative attack. We run 40 iterations of step size $2/255$ under $\epsilon = 8/255$. We implement three variants: (i) naive PGD, which attacks the classification model normally (ignoring the defense); (ii) semantic PGD (S-PGD), which constrains perturbations to the object region only; and (iii) transformation-invariant PGD (S-TI-PGD), which applies random affine transformations during the attack. S-TI-PGD is conceptually similar to Expectation-over-Transformation attacks by [15] aimed at defeating defenses with random input transformations.

For adversarial patch attacks, we use a setup similar to, but not quite the same as the original paper on adversarial patches [16]. We allow the attacker to place a square patch (of size 64×64 pixels in our experiments, roughly 5–10% of the image area) anywhere in the image with arbitrary pixel values. The patch is optimized using 100 steps of PGD on the patch pixels to maximize the classifier's loss. We consider two scenarios: a naive patch that can be placed anywhere and a semantic patch (S-Patch) that is restricted to cover the object of interest (as a worst-case for 3DPure, since a patch on the background might be removed entirely by segmentation).

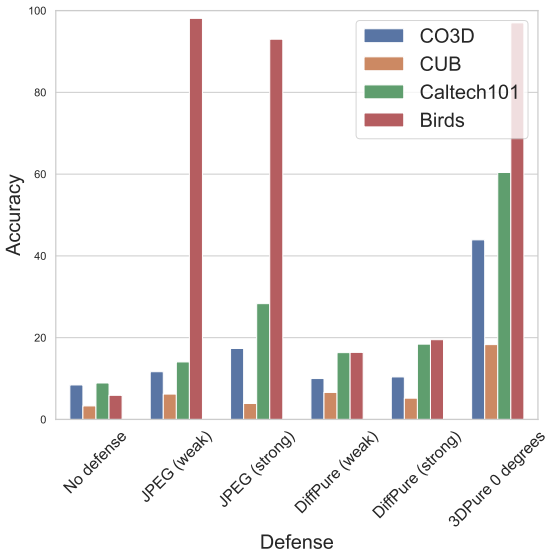

Fig. 2. Robust accuracy comparison of baseline purification methods and 3DPure against Semantic Patch Attack.

Baseline Defenses: We compare 3DPure against two representative input purification defenses: (1) JPEG Compression—a simple but common defense that compresses the image and decompresses it, which can nullify high-frequency perturbations. We evaluate two levels: JPEG (weak) with quality 25, and JPEG (strong) with quality 5. (2) DiffPure—the diffusion-based purification method by [4]. A pretrained diffusion model is used with two noise levels, $\sigma = 0.1$ and $\sigma = 0.3$, analogous to weak and strong purification respectively. (3) We also consider the baseline of no defense, to gauge the original accuracy and vulnerability of CLIP.

We measure two metrics for each method: clean accuracy (accuracy on unperturbed test images) and robust accuracy (accuracy on adversarially attacked images). These are reported for each dataset and attack type, following standard evaluation protocols.

V. Results and Discussion

To start, we assess how 3DPure affects clean accuracy, as shown in Table I. As expected, applying any input transformation tends to reduce the accuracy of the classifier due to image distortion. 3DPure achieves a clean accuracy that is slightly lower than most baselines, and comparable to the strong JPEG defense. For instance, on the CO3D dataset, CLIP's clean accuracy is 80.6% with no defense, which drops to 79.8%

TABLE I
COMPREHENSIVE ACCURACY RESULTS (%).

Dataset	Method	No Attack	S-FGSM	PGD	S-PGD	S-TI-PGD	Patch Attack	S-Patch Attack
CO3D	No Defense	**80.6**	53.6	3.6	12.2	15.6	1.3	8.4
	JPEG (weak / strong)	79.8 / 68.2	**62.1** / 39.0	39.9 / 38.0	41.8 / 37.5	22.9 / 34.4	4.0 / 10.7	11.7 / 17.4
	DiffPure (weak / strong)	71.0 / 64.2	58.1 / 56.6	45.4 / **59.3**	47.2 / **54.2**	20.4 / **38.0**	3.7 / 3.8	10.0 / 10.4
	3DPure (ours)	62.9	56.1	43.0	47.6	28.3	**18.7**	**43.9**
CUB	No Defense	**64.2**	35.6	2.3	10.8	13.9	0.3	3.3
	JPEG (weak / strong)	62.5 / 41.5	**40.1** / 15.1	20.5 / 22.2	27.9 / 14.6	18.4 / 13.3	0.9 / 2.5	6.2 / 3.9
	DiffPure (weak / strong)	56.0 / 46.1	39.0 / 36.0	23.5 / **34.4**	32.5 / **32.7**	21.4 / **23.9**	1.4 / 1.4	6.6 / 5.2
	3DPure (ours)	41.0	34.6	25.4	28.5	16.8	**6.7**	**18.3**
Caltech101	No Defense	**90.3**	67.2	11.3	20.0	34.3	2.0	8.9
	JPEG (weak / strong)	89.9 / 74.9	**73.6** / 56.7	53.3 / 67.0	56.4 / 55.0	44.7 / 51.6	7.4 / **34.6**	14.1 / 28.3
	DiffPure (weak / strong)	89.6 / 86.3	73.1 / 73.4	57.6 / **71.7**	61.4 / **68.9**	42.2 / **53.9**	9.7 / 12.0	16.3 / 18.4
	3DPure (ours)	75.7	72.4	56.5	65.7	51.0	25.6	**60.4**
Birds	No Defense	**100.0**	97.3	12.1	54.9	67.5	0.5	5.9
	JPEG (weak / strong)	100.0 / 99.7	98.1 / 93.0	**98.1** / 93.0	**98.1** / 93.0	**98.1** / 93.0	**98.1** / 93.0	**98.1** / 93.0
	DiffPure (weak / strong)	99.8 / 99.5	98.8 / **98.9**	92.2 / 96.6	95.9 / 97.0	79.5 / 86.0	5.0 / 4.7	16.4 / 19.5
	3DPure (ours)	99.5	98.0	97.3	97.1	87.3	97.1	97.0

with a weak JPEG, 71.0% with DiffPure, and down to 62.9% with 3DPure (frontal view). The drop is more pronounced on CUB (e.g. 3DPure yields 41.0% vs. 64.2% no defense), likely because the detailed parts of the bird species become unrecognizable after reconstruction.

Next, we evaluate robust accuracy under adversarial attacks. Table I presents a detailed account of the results, and we highlight key findings below.

Patch Attacks: 3DPure excels at defending against adversarial patches. In Figure 2, we see that without any defense, the semantic patch attack completely breaks CLIP—robust accuracy is near 0% on all datasets. JPEG compression offers some help, but even the strong JPEG defense yields $< 30\%$ accuracy in all cases except the Birds dataset. DiffPure, which performed well against distributed noise, struggles a lot. Both weak and strong DiffPure versions obtain under 20% robust accuracy on all datasets. In stark contrast, 3DPure raises the robust accuracy to 40–60% on the CO3D and Caltech101 datasets. The superiority of 3DPure on patch-based attacks is largely due to the 3D reconstruction step removing the patch from the scene: since the patch is an arbitrary flat pattern not consistent with the object's true appearance, the reconstructor either ignores it or severely distorts it in the process of modeling the object. As seen in Figure 3, the rendered images from 3DPure show that the adversarial patch is broken apart and does not appear as a coherent, effective pattern to the classifier. These results confirm 3DPure's strength for localized attacks.

FGSM Attack: FGSM is a relatively weak attack, and interestingly, we find that CLIP without defense is already somewhat robust to it. CLIP's accuracy drops by only a few points under S-FGSM. As a result, all defenses appear to yield only minor gains, and in some cases, they hurt. On CUB, we observed that 3DPure actually performed slightly worse than no defense for S-FGSM, bringing accuracy from 35.6% down to 34.6%. This indicates that for low-strength attacks, the reconstruction-induced distortion may do more harm than the perturbation itself. In summary, S-FGSM did not pose a serious challenge for CLIP, so the benefits of 3DPure are not pronounced in this case. This reinforces the notion that one might not want to deploy a heavy defense like 3DPure unless expecting stronger attacks.

PGD Attack: Without any defense applied, all PGD variants dramatically reduce CLIP's accuracy, dropping it to approximately 10–30% on datasets such as Caltech101. The S-TI-PGD variant has the lowest attack success rate but is particularly challenging to purify, highlighting how the Expectation-over-Transformation (EoT) mechanism strengthens attacks against various defenses. DiffPure consistently provides the strongest improvement in robust accuracy, while JPEG-based defenses offer limited benefits. 3DPure achieves robust accuracy close to DiffPure's in most scenarios and, notably, nearly matches DiffPure's performance on the Birds dataset (excluding S-TI-PGD). This similarity is likely due to Birds being a low-detail dataset that tolerates significant perturbations. Overall, these findings show that 3DPure maintains moderate effectiveness even against spatially distributed adversarial attacks.

Viewpoint Jitter Analysis: During development, we tested the effect of using alternate viewpoints in the rendering step (as an additional defense layer). We rotated the camera around the object by various angles and measured CLIP's accuracy on clean images (Fig. 4). Consistent with intuition, the frontal view (0° rotation, the original viewpoint) gave the best accuracy in general. Side angles (90°) often led to missing object parts (e.g. seeing a bird from the side or back makes species identification harder) and thus lower accuracy. Interestingly, for symmetric objects or when the back view is similar to the front, accuracy had a secondary peak. Overall, we found that jittering the view did not provide a net benefit under attack: while it can confuse an attacker who does not anticipate it, it also confuses the classifier if the angle is suboptimal.

Discussion: The experimental results confirm that 3DPure

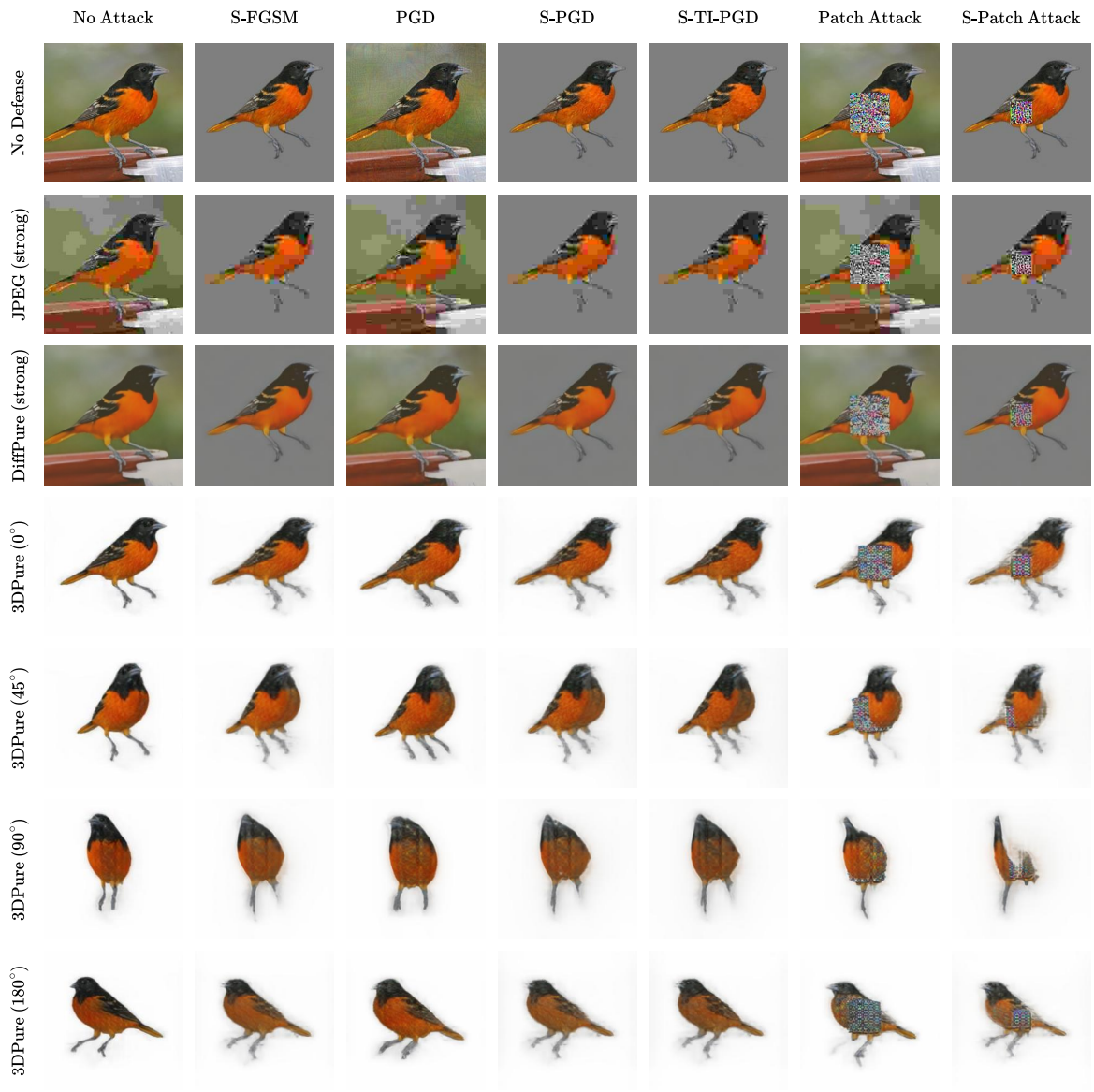

Fig. 3. Visualization of defense effectiveness across various attacks and purification methods.

is especially powerful against localized adversarial noise like patches. By reconstructing the object geometry, 3DPure essentially "forgets" the adversarial patch—the patch either is not reconstructed at all or appears as an unnatural artifact that does not resemble the adversarial pattern from the original view, hence failing to mislead the classifier. Against broader ℓ_∞ attacks, 3DPure performs almost as well as state-of-the-art 2D purification (DiffPure), demonstrating that even those pixel-wise perturbations cannot fully survive the 3D transformation. One caveat observed is the reduced clean accuracy and slight performance dip on high-detail classification tasks, due to the inherent information loss in modeling and re-rendering. This suggests 3DPure is most suited for scenarios where adversarial attacks are a real concern and a small hit to clean performance is acceptable.

VI. Conclusion

We presented 3DPure, a new approach to adversarial defense that incorporates single-image 3D reconstruction into the classification pipeline. By lifting images into a 3D representation and reprojecting them, 3DPure enforces a form of consistency that eliminates adversarial perturbations misaligned with physical reality. Our implementation using TripoSR and CLIP demonstrates that this concept is not only feasible but highly effective: 3DPure significantly improves robust accuracy under strong adversarial attacks, especially completely nullifying

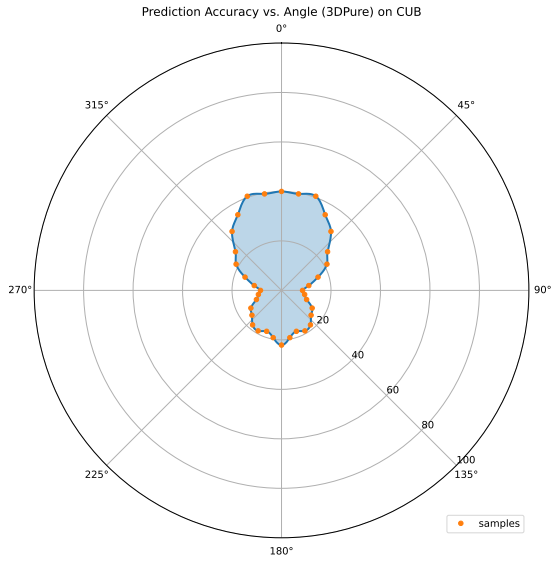

Fig. 4. Clean accuracy of CLIP under camera rotation. Orange dots represent individual samples, while the blue curve shows interpolated values.

patch attacks that defeat conventional defenses. It does so while maintaining competitive performance against standard ℓ_∞ attacks and incurring a reasonable trade-off in clean accuracy.

This work opens several directions for future research. Evaluating 3DPure on mixed datasets with both clean and adversarial inputs—augmented with a detection module—would better reflect real-world deployment scenarios. Case studies targeting practical tasks, such as secure product image verification, could further demonstrate its applicability. To probe its limits, stronger attacks incorporating Expectation-over-Transformation (EoT) could be used to generate perturbations that survive the reconstruction pipeline. Exploring universal perturbations transferable across different 3D methods, and analyzing robustness under variations in viewpoint and lighting, would deepen understanding of its resilience. Finally, testing adaptive attacks—particularly those estimating gradients through 3DPure—remains critical for assessing true robustness under adversarial pressure.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grants JP21H04907 and JP24H00732, by JST CREST Grant JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, and by JST K Program Grant JPMJKP24C2 Japan.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[3] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[4] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, *Diffusion models for adversarial purification*, 2022. arXiv: 2205.07460.

[5] Z. Wu, Y. Duan, H. Wang, Q. Fan, and L. J. Guibas, *IF-Defense: 3d adversarial point cloud defense via implicit function based restoration*, 2020. arXiv: 2010.05272.

[6] C.-H. Ho and N. Vasconcelos, "DISCO: Adversarial defense with local implicit functions," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, Curran Associates, Inc., 2022, pp. 11 731–11 744.

[7] B. Liu, J. Ji, C. Tao, J. Li, and Y. Wang, "The adversarial robust and generalizable stereo matching for infrared binocular based on deep learning," *J. Imaging*, vol. 10, no. 11, p. 264, 2024.

[8] L. Li, Q. Lian, and Y.-C. Chen, "Adv3D: Generating 3d adversarial examples for 3d object detection in driving scenarios with NeRF," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[9] X. Yuan and I. K. Park, "Face de-occlusion using 3d morphable model and generative adversarial network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10 061–10 070.

[10] D. Tochilkin, D. Pankratz, Z. Liu, *et al.*, *TripoSR: Fast 3d object reconstruction from a single image*, 2024. arXiv: 2403.02151.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, PMLR, 2021, pp. 8748–8763.

[12] P. Welinder, S. Branson, T. Mita, *et al.*, "Caltech-ucsd birds 200," California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010.

[13] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10 901–10 911.

[14] F.-F. Li, M. Andreeto, M. Ranzato, and P. Perona, *Caltech 101*, 2022.

[15] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 80, PMLR, 2018, pp. 274–283.

[16] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, *Adversarial patch*, 2017. arXiv: 1712.09665.