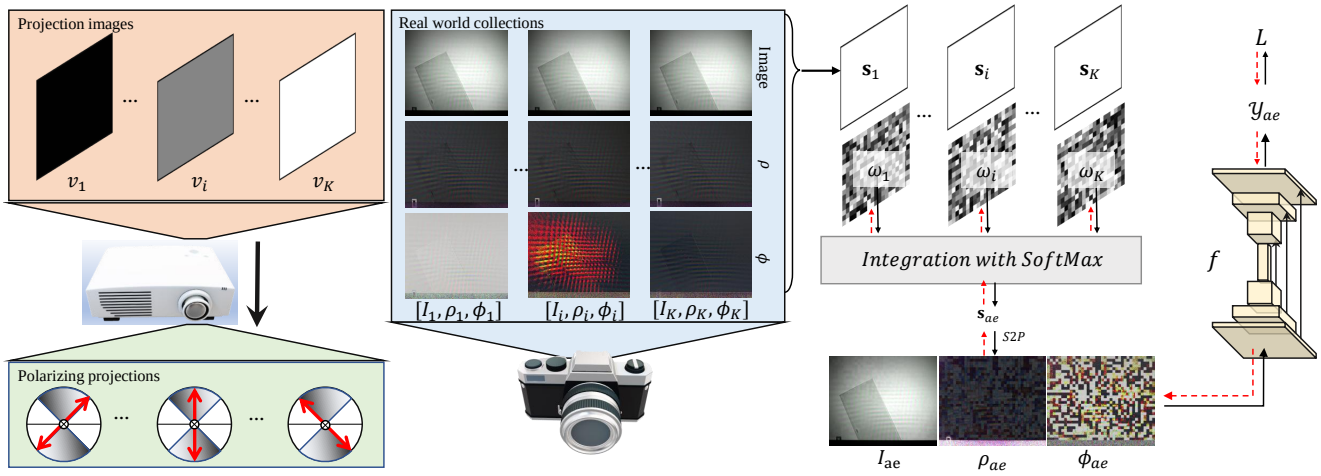
Fooling Polarization-based Vision using Locally Controllable Polarizing Projection Supplementary Material

Zhuoxiao Li[1] Zhihang Zhong[2] Shohei Nobuhara[3] Ko Nishino[3] Yinqiang Zheng[1*]
[1]The University of Tokyo [2]Shanghai Artificial Intelligence Laboratory [3]Kyoto University



(a) Projection (b) Basis images (c) Adv. Example Generation (d) PGSNet

Figure 1. Illustrations of (a) polarizing projections, (b) candidate images collection, (c) adversarial example generation, and (d) feed forward gradient backward of the polarization-RGB-based glass segmentation model PGSNet [2]. v_i represents the pixel value for the projected monochromatic image. $[I_i, \rho_i, \phi_i]$ denote captured polarization cues, including intensity, DoLP, and AoLP maps, and s_i are the same data in the form of Stokes parameters. S2P represents the conversion between Stokes parameters and Polarization cues. Black and red dashed lines in (c), (d) counts for flows of data feed-forward and gradients backward propagation.

1. Whitebox Attack on Glass Segmentation

1.1. Implementation Details

We propose to employ locally controllable polarizing projection to fool the polarization-RGB-based glass segmentation model, PGSNet [2], in the physical world. We control the polarizing projection by projecting specific images according to the principle of our adapted one-chip LCD projector. In general, physical world adversarial attacks require realistic simulations of targeted objects as well as scenes injected with adversarial perturbations, e.g., projection, shadow, and stickers, to search a robust and effective adversarial perturbation pattern. However, polarization reflection modeling requires accurate pBRDF parameters and geometries that are unavailable in the wild. Thus, we propose a simplified approach for constructing a simulation of high precision. Our perturbation is a map of grids, which optimized to construct a perturbation robust enough to be captured in the complex real world. The adversarial perturbation should be estimated in an optimization-based approach in whitebox attack. To avoid physics-based simulation of polarized light transports from the projector to the camera, which involves the non-linear function of the projector and polarization reflection modeling, we generate adversarial examples from a set of candidate real captures whose projection values are known. Concretely, we first set the projector (ELEPHAS W13 after hardware modification) and RGB-P camera (BFS-U3-51S5PC-C equipped with the IMX250MYR Polar-RGB sensor) in the same place and both look forward to the target scene.

*Corresponding author

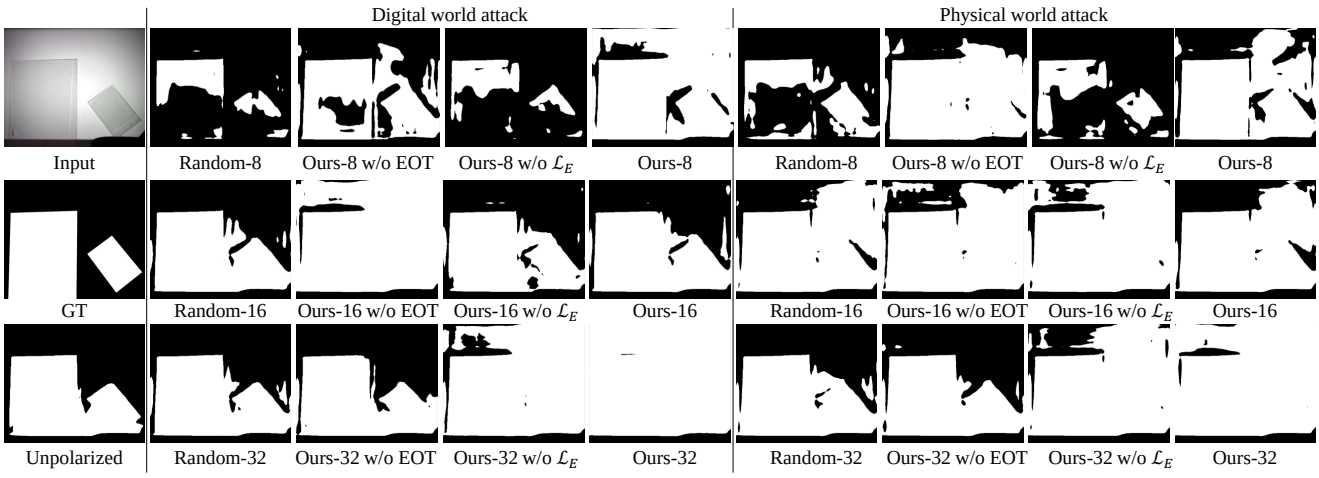


Figure 2. Visual comparisons for adversarial attacks on the Deep SfP-wild [2].

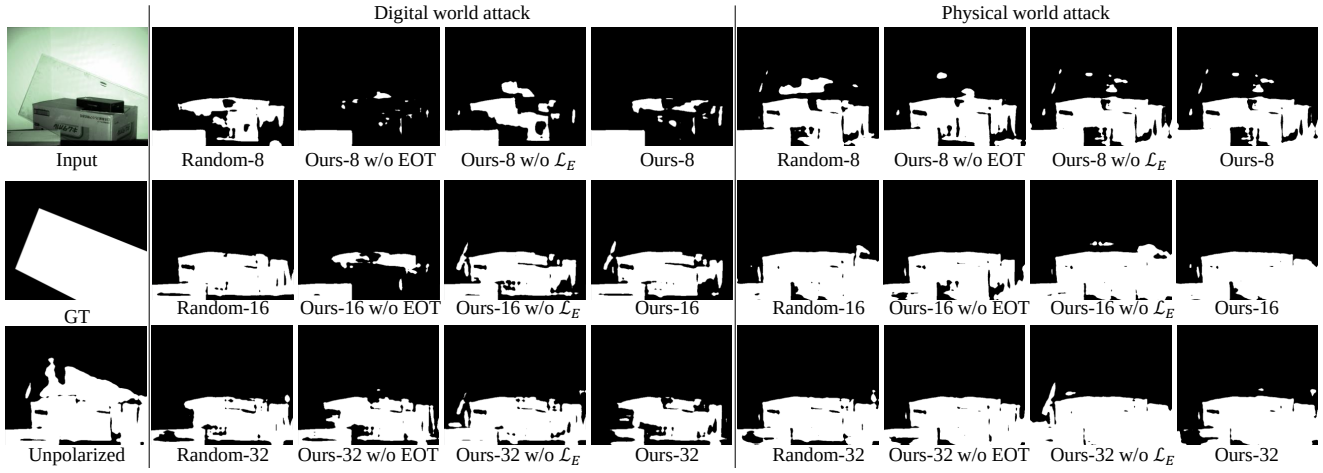


Figure 3. Visual comparisons for adversarial attacks on the Deep SfP-wild [2].

Then, we project flat grayscale images of a gray level $\mathbf{v}_p = \{v_1, v_2, ..., v_K\}$ uniformly sampled from 0 to 255, which causes projections with constant polarizing directions, as shown in Figure 1(a). Next, we capture a sequence of real world images as a set of basis images. As shown in Figure 1 in the form of polarization cues, in contrast to few differences between the intensity images I, their AoLP maps ϕ are significantly changed by polarizing projections. In Figure 1(c), the basis images in the form of Stokes parameters are integrated into sae with a set of optimizable weight maps $\mathbf{\Omega} = \{\omega_1, \omega_2, ..., \omega_K\}$ through the $SoftMax$ function as:

$$\mathbf{s}_{ae} = \sum_{i}^{K} \frac{\exp(\omega_i/\tau)}{\sum_j^K \exp(\omega_j/\tau)} (\mathbf{s}_i - \mathbf{s}_b) + \mathbf{s}_b^*. \qquad (1)$$

The integration $[I_{ae}, \rho_{ae}, \phi_{ae}]$ is fed into PGSNet as an adversarial example, to optimize the coefficient maps Ω, as shown in Figure 1(d). After the optimization, the best projection value in each grid is selected from $\{v_1, ..., v_K\}$ by $ArgMax$ function.

1.2. Experiment

Figure 2 and 3 present additional visual results from our polarization projection attacks. When contrasted with results under unpolarized illumination or those stemming from random perturbations, it becomes evident that our optimized polarizing projections consistently effectuate potent and resilient adversarial attacks in real-world settings. Figure 3 illustrate the results of glass detection and attacks against a complex background, indicating that the model has some ability to recognize complex backgrounds but is more sensitive to attacks. Therefore, even random projections can affect accuracy, which proves the effectiveness of our proposed polarization projection device in attacking

polarization-based vision models.

2. Whitebox Attack on SfP-wild

2.1. Implementation Details

We employ an identical approach when evaluating our attacks on the state-of-the-art Deep SfP-wild model [1]. Specifically, we establish a grid size of 2. The Gaussian blur is applied with a kernel size of 7×7, and the standard deviation is sampled within the range of 4 to 6. Moreover, we designate a step size of $\alpha = 1000$ and undertake 500 iterations for perturbation updates.

2.2. Experiment

Figure 4 presents an extended set of visual outcomes from our physical world attacks. Implementing such an adversarial attack is challenging due to several factors: the markedly low precision of projection, the disparity between digital and physical realms, and the constraints of available polarization reflection patterns. Despite these hurdles, our results compellingly demonstrate that our approach can effectively deceive the Deep SfP model. The quantitative comparison for the collected 2 scenes are shown in Table 1. A perturbation with a lower resolution and optimized with EOT takes more robust attakcing performance in a physical-world scenario.

Table 1. Quantitative comparison. The numbers denote grid sizes, and * represents the perturbation generated w/o EOT.

MAE (°) ↓	Ours-1*	Ours-2*	Ours-1	Ours-2
Digital	25.76	26.50	21.11	23.08
Physical	43.64	41.73	41.02	**38.23**

3. Fooling DoLP-based Color Constancy

The DoLP-based color constancy algorithm [3] relies on polarization properties of reflections to infer the global illumination. The algorithm uses DoLP to search achromatic pixels within an image, whose reflection color directly reflects color of light source. For chromatic pixels, a robust prior knowledge and an assumption are applied, that their white-balanced color should be opposite to its DoLP color, and the AoLP of secular and diffuse reflection polarizing in orthogonal directions. For most lighting conditions in the wild, the assumption is well grounded as illuminations of non-polarization or low-polarization dominate, and global illuminations generally convey limited intensity caused by reflections from other objects or interreflections. However, our polarizing projection is capable of emitting linearly polarized light through separate color channels.

As the algorithm utilizes simple statistic computation to estimate a global lighting ratio, we believe an intuitive yet effective way can destroy the color constancy algorithm, by simply projecting colors of three channels in different polarizing directions. To evaluate it, we apply several simple projection settings. While diffuse polarization is barely influenced by polarization state of incident light, specular reflections well retain linearity of incident light.

3.1. Experiment

Figure 5 showcases images that have been restored using estimated white-balance ratios under conditions heavily influenced by our polarizing projections. Notably, these images are predominantly characterized by specular reflections. Intriguingly, the results reveal that despite reflections exhibiting elevated DoLP values in certain channels (such as blue and green), the algorithm is misdirected to estimate a cyan illumination. This misestimation further amplifies the red channel, as evidenced in Figure 5 by the RGB value [255, 0, 0].

4. Blackbox Attack on HPSfP

The human pose and shape estimation from polarization (HPSfP) [4] method proposes to employ the polarization cues for the specific shape estimation scenario, human, and further leaveraging the geometric cues for human pose estimation. This approach significantly outperforms prior RGB-based methods. However, our experiments reveal its performance instability under polarized projections. As illustrated in Figure 6, while the method estimates resonable human pose and shape under unpolarized illumination y_{up}, estimations under uniformly polarized projection $y_{ae}^{45^\circ}$ and randomly polarized projections of varying resolutions y_{ae}^{k} significantly diverge from the original results.

References

[1] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12632–12641, 2022. 3, 4

[2] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12622–12631, 2022. 1, 2

[3] Taishi Ono, Yuhi Kondo, Legong Sun, Teppei Kurita, and Yusuke Moriuchi. Degree-of-linear-polarization-based color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19740–19749. IEEE, 2022. 3, 4

[4] Shihao Zou, Xinxin Zuo, Sen Wang, Yiming Qian, Chuan Guo, and Li Cheng. Human pose and shape estimation from single polarization images. *IEEE Transactions on Multimedia*, 2022. 3, 5

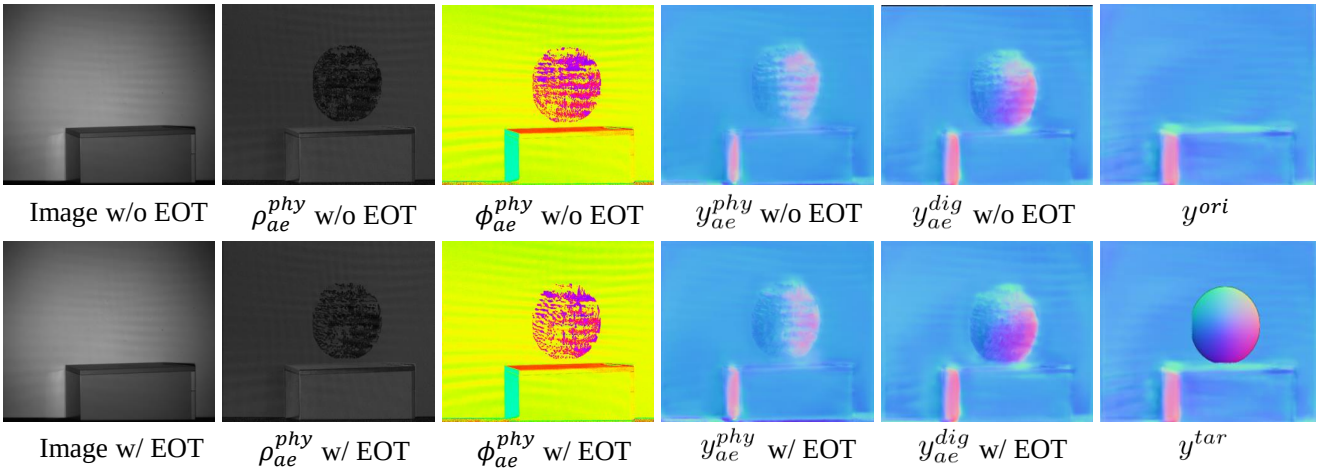


Figure 4. Visual comparisons for adversarial attacks on the Deep SfP-wild [1].

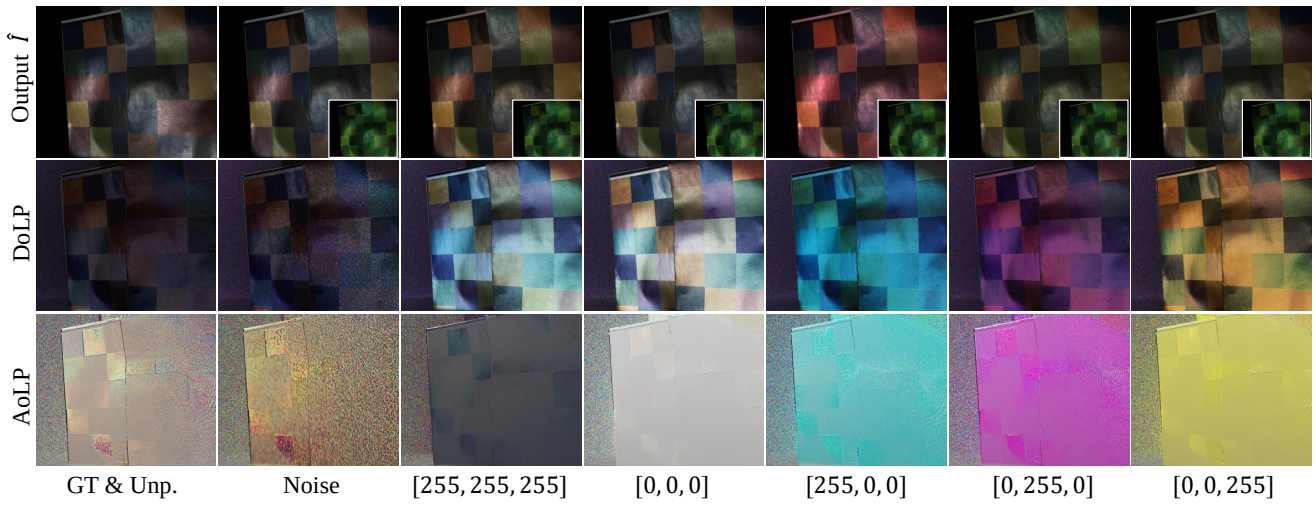


Figure 5. Color constancy [3] results and visualizations of DoLP, AoLP under different color-wise polarization projections. Notice that our projection is always constant white light, and the labels denote color-wise values of polarizing perterbation.

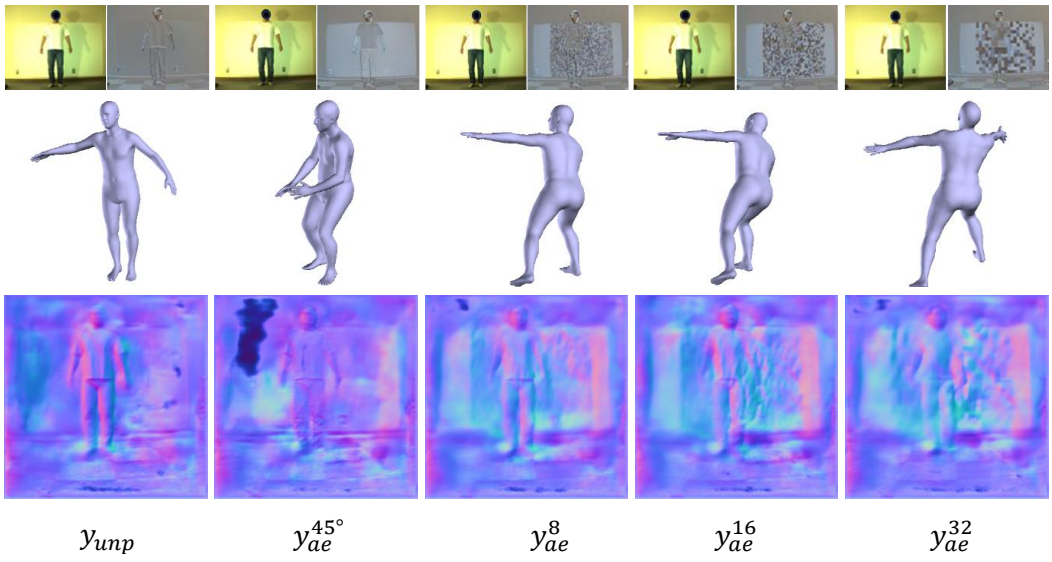


Figure 6. Illustrations for blackbox attacks on HPSfP [4] using different projections. The figures display both the images and AoLP maps alongside the resulting pose and shape estimations.