Fooling Polarization-based Vision using Locally Controllable Polarizing Projection

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

Abstract

Polarization is a fundamental property of light that encodes abundant information regarding surface shape, material, illumination and viewing geometry. The computer vision community has witnessed a blossom of polarization-based vision applications, such as reflection removal, shape-from-polarization (SfP), transparent object segmentation and color constancy, partially due to the emergence of single-chip mono/color polarization sensors that make polarization data acquisition easier than ever. However, is polarization-based vision vulnerable to adversarial attacks? If so, is that possible to realize these adversarial attacks in the physical world, without being perceived by human eyes? In this paper, we warn the community of the vulnerability of polarization-based vision, which can be more serious than RGB-based vision. By adapting a commercial LCD projector, we achieve locally controllable polarizing projection, which is successfully utilized to fool state-of-the-art polarization-based vision algorithms for glass segmentation and SfP. Compared with existing physical attacks on RGB-based vision, which always suffer from the trade-off between attack efficacy and eye conceivability, the adversarial attackers based on polarizing projection are contact-free and visually imperceptible, since naked human eyes can rarely perceive the difference of viciously manipulated polarizing light and ordinary illumination. This poses unprecedented risks on polarization-based vision, for which due attentions should be paid and counter measures be considered.

1. Introduction

Even if the frequency of light lies in the visible range, its polarization status can hardly be perceived by human eyes. Fortunately, a variety of imaging devices have been developed, which allow to utilize rich scene information encoded in polarization, regarding geometry, material, illumination and light transportation. The emergence of single-chip mono/color polarization sensors has made polarization data acquisition easier, leading to a blossom of polarization-based vision applications, such as reflection removal [22], shape-from-polarization (SfP) [10, 13, 23], surface defects detection [24], color constancy [28], transparent object detection and segmentation [26]. Figure 1 (a) shows the capabilities of a single-chip color polarization sensor in capturing trichromatic image I, angle of linear polarization (AoLP) ϕ, and degree of linear polarization (DoLP) ρ, with one shot. Given the prevalence of polarization-based vision, it is astonishing that its vulnerability has never been formally explored in the CV and AI communities.

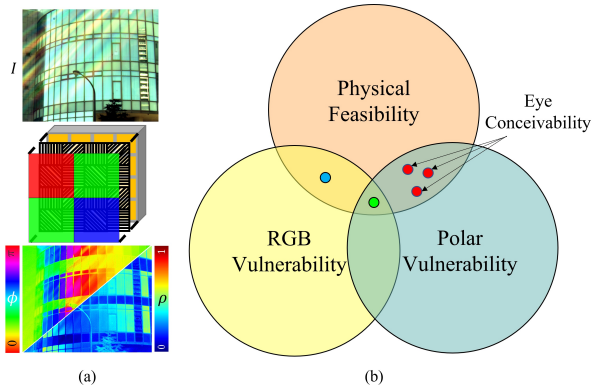


Figure 1. (a) Single-chip color polarization sensor can capture trichromatic image, angle of linear polarization (AoLP), and degree of linear polarization (DoLP) within one shot. (b) Our proposed physical attackers are based on polarizing projection, which is naturally conceivable to human eyes, thus can bypass the trade-off between attack efficacy and eye conceivability in fooling RGB-based vision.

The vulnerability of RGB-based deep vision models is firstly reported in [30], with various extensions in minimizing perturbation magnitude, maximizing success rate of attack, retrieving universal adversarial attackers, and so on [1, 2]. Stepping beyond the digital space, more recent researches focus on studying the vulnerability of RGB-based vision models in the physically feasible space, while minimizing the level of offensiveness to human eyes (blue

*Corresponding author

point in Figure 1(b)), by using printed attack patterns on papers [5, 19, 21], clothes [34], or perturbations projected by projectors [14], and attacks created by laser beams [11] and shadows [38]. In principle, physical adversarial attack can be more catastrophic, since there is no need to hack the input image or the deployed model as digital attack requires. However, it is extremely hard to find a physical attacker that is both effective and imperceptible, since RGB cameras by design are mimicking human eyes, and a physically feasible attack that is invisible to eyes will not be captured by the camera as well.

Given that polarization and color represent two distinct dimensions of light, and polarization is usually introduced as a complementary modality to assist RGB-based vision, one might believe that, polarization-based vision, especially when coupled with the RGB modality, should be safer and harder to be attacked in the physical world, as the green dot in Figure 1(b) illustrates. In this paper, we show that this speculation is ungrounded by proposing a novel yet simple implementation of locally controllable polarizing projection. Since human eyes have no sensitivity to polarization, the most stringent restriction on eye conceivability in attacking RGB-based vision is naturally bypassed. This allows us to explore the vulnerability space of polarization-based vision more flexibly, within in the broad feasible space that the projector can realize (red dots in Figure 1(b)).

Inspired by the operating principle of Liquid Crystal Display (LCD) panels in monitors and projectors, we have recognized that the polarization status of light emitted from each liquid crystal cell can be independently controlled, after removing the front polarization film attached onto the LCD panel. Since human eyes can not perceive polarization status, the projected light looks uniformly white, even if the projection pattern has colors and textures, and the polarization status of light has been adjusted accordingly by LCD. In contrast, polarization cameras can record the programmed polarizing projection, and the behaviors of vision algorithms based on such information might be manipulated.

We have verified the feasibility of fooling polarization-based vision for two representative tasks via whitebox attack, including (i) reducing the accuracy of RGB-polar-based glass segmentation [26]; (ii) misleading the latest shape estimation model on the basis of polarization [23]. We hope this study can arouse attention on the potential security risks of utilizing polarization and trigger further researches on the defense side.

2. Related Work

2.1. Adversarial Attacks on RGB-based Vision

While the state-of-the-art deep neural networks are capable of achieving incredible performance in various scene understanding tasks, recent researches [21, 30] revealed their striking vulnerability that very mild modifications to the input images can deceive advanced classifiers with high confidence. The adversarial examples are generated through optimization processes by maximizing the classification error of a targeted model. In digital world, on the premise of direct access to the targeted model, an adversarial example can be derived by one or multiple steps of perturbation following negative gradient directions, including classic Fast Gradient Sign Method (FGSM) [15], the Basic Iterative Method (BIM) [15], and the Projected Gradient Descent (PGD) [25] for efficient and transferable adversarial attacks. Their perturbations are bounded with a small norm-ball $L_p < \epsilon$, normally $p = 2$ or ∞, or minimized with a joint adversarial loss [8], to craft a quasi-imperceptible example to human eyes.

Digital attacks assume they can hijack the prediction system to directly feed adversarial examples into the targeted model. Considering that this requirement is usually impractical, other researches try to realize adversarial attacks by inserting perturbations into the physical world. [21] shows adversarial examples printed on papers are partly effective to fool DNN classifiers. However, because of the discrepancy between the designed attacker in the digital space and the physical attacker recorded by the camera, a key task is to retrieve robust adversarial examples that can be faithfully realized. [19] approximates the full digital-to-physical transformation to search perturbations in a simulated world. To deal with the wide range of diversities in real world scenarios, e.g. view points, illuminations, and noises, [5] gets a distribution of transformations involved in the optimization procedure, including rescaling, rotation (in 2D or 3D), translation of image, and so on.

However, former small perturbations are too subtle to be captured by cameras in the wild. Therefore, recent physical-world adversarial attack methods attempt to generate strong but stealthy perturbations in the real world. For example, stickers and graffiti-type perturbations are attached to targeted objects, e.g. a road sign, to achieve targeted misclassification from arbitrary viewpoints. Wearable attack perturbations like clothes [35] and eye-glasses [29] are capable of fooling detection systems with improved stealthiness. Moreover, laser beams [11], shadows [38], and projection [14] are utilized to craft attack perturbations in the physical world without touching target objects. We refer readers to [33] for thorough literature reviews on physical adversarial attacks. All these researches on physical adversarial attacks have to make a trade-off between attack efficacy and eye conceivability.

2.2. Polarization-based Vision

Polarization has been utilized in various vision tasks for many years, which is further boosted recently due to the

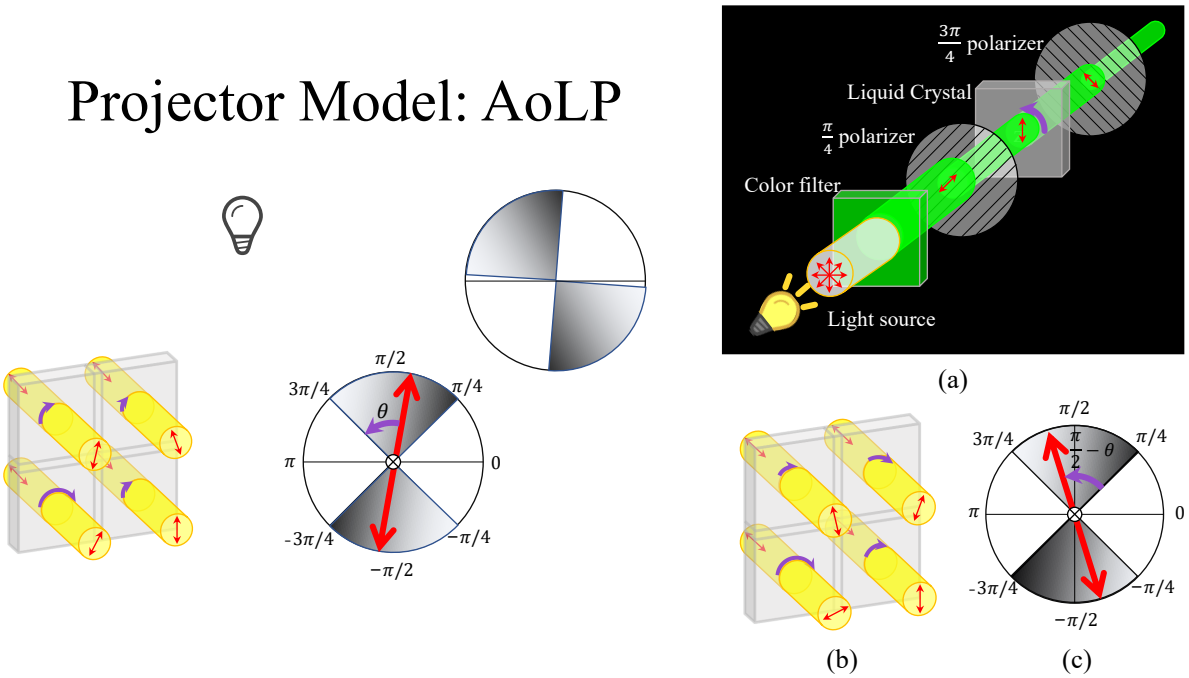


Figure 2. (a) The mechanism of intensity adjustment in one-chip LCD projector with a liquid crystal panel sandwiched by two perpendicular linear polarizers. (b) The polarization direction of the light beam in each liquid crystal cell can be individually controlled, without affecting its intensity. (c) The range of controllable polarization direction.

wider applications of polarization in the near future can be expected, yet we would like to warn of the potential vulnerability of polarization-based vision, which might be more serious than that of RGB-based vision, since the adversarial attackers can be physically realized using a modified LCD projector and human eyes can not differentiate maliciously manipulated polarizing light from normal illumination.

2.3. Projectors and Their Applications

Projectors are widespread display devices, whose modulation mechanism of light intensity is either based on digital micromirror device (DMD) widely used in digital light processing (DLP) projectors or liquid crystal polarization adopted by LCD projectors or LCoS projectors. As for the color-framing mechanism, one-chip DLP projectors use the rotating color wheel or blinking trichromatic LEDs, and one-chip LCD projectors use the micro color filter array, which is similar to the Bayer pattern in RGB cameras. By using the color-framing mechanism of a one-chip DLP projector, Ashdown et al. [4] recovered high-resolution spectral reflectance. Further, to deal with unexpected irregularities when applying a digital projector in non-ideal situations, they proposed to generate a compensation image based on both the radiometric model of the system and the content of the image. Tanaka et al. [31] utilized a projector coaxially placed with the camera to inject illuminations of multiple frequencies for obtaining the appearance of individual inner slices. Projectors have also been used for adversarial attacks in the physical world with projected perturbations [14] or constant colors [17].

Existing polarization-based LCD/LCoS projectors do not offer pixel-wise manipulation of the polarization status of light projected on the screen. So, they can not be directly utilized to attack polarization-based vision algorithms. In the following, we will show that a simple adaptation of the one-chip LCD projector will allow locally controllable polarizing projection.

3. Preliminary

3.1. Principle of One-chip LCD Projector

One-chip LCD projector is the most widely used type of low-cost projector. The principle of an LCD projector controlling the irradiance is shown in Figure 2. A beam of unpolarized light is emitted by a bulb. Two linear polarizers are placed in coaxial positions while their polarizing directions are perpendicular to each other ($\frac{\pi}{4}$ and $\frac{3\pi}{4}$ in our device), and a liquid crystal panel is inserted to the middle of them. The light is divided into red, green, and blue components by a color filter array before linearly polarized by the back $\frac{\pi}{4}$ polarizer. Then, by adding voltages to liquid crystal grids, the layer can manipulate polarizing direction of individual light beams, and a greater voltage leads to a bigger rotation up to $\frac{\pi}{2}$ from its initial direction. The light

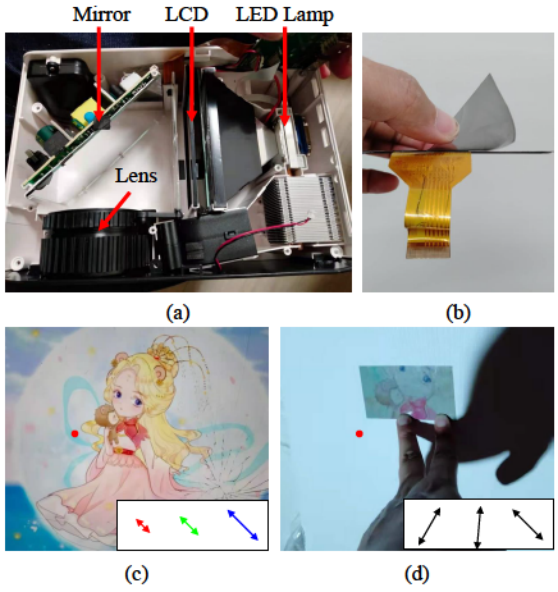


Figure 3. (a) The typical structure of a one-chip LCD projector, in which light emitted by LED lamp will go through a liquid crystal panel, mirror, and projection lens. (b) A linear polarizer is attached to the front side of the LCD panel. We tear it off to make our polarizing projection. (c)(d) The projection of a normal LCD projector and our adapted projector. The arrows' direction and length represent light polarizing direction and intensity, respectively. A normal projector emits out colorful light of constant polarization direction, while our adapted polarizing projector emits light with constant intensity but different polarizing angles. Note that, for naked eyes and ordinary RGB cameras, the projected light is completely uniform, even if their polarization directions are totally different. The projected image can be observed by eyes with the assistance of a linear polarizer on the screen.

intensity passing through the front polarizer is decided by its polarizing direction, following the Malus's law:

$$I = I_0 \cos^2(\theta), \tag{1}$$

where θ is the angle between the polarizing direction of light after being rotated by the liquid crystal and the direction of the front $\frac{3\pi}{4}$ polarizer. I_0 is a constant light intensity from the back polarizer. Since the intensity of RGB components is separately controlled, the color of merged light can be manipulated to match the projection pattern sufficiently.

Note that, with the front polarizer equipped, output light beams are always linearly polarized in the $\frac{\pi}{4}$ axis related to the projector but have different intensities, as shown in Figure 3 (c). *The key idea of building a controllable polarization light projector is to remove the front polarizer.* As shown in Figure 3 (a,b), we tear off the front polarization film of the projector and manage not to damage the liquid crystal panel. In this way, the output lights of the projector have constant intensity but different polarizing directions. The polarizing direction can be precisely controlled by manipulating the projection pattern. Furthermore, uniform white color and constant projection intensity contribute to high stealthiness as it will not introduce visible textures to human eyes, as can be seen in Figure 3 (d).

3.2. Preliminaries for Light Polarization

Most polar-RGB based methods rely on both intensity and polarization cues, i.e., degree of linear polarization (DoLP, ρ, the proportion of linear polarized component in light) and angle of linear polarization (AoLP, ϕ, the polarizing direction of polarized light). They can be calculated from a single shot with a Bayer-polarization sensor, *e.g.*, IMX250MYR, which captures polarization components in four directions, termed as I_0, $I_{\frac{\pi}{4}}$, $I_{\frac{\pi}{2}}$, and $I_{\frac{3\pi}{4}}$. Stokes parameter, $\mathrm{s} = [s_0, s_1, s_2]^\top$ is used to describe the polarization state of light, where s_0 represents the total intensity of light, s_1 and s_2 describe the polarization states in horizontal and diagonal axes. s_0, s_1 and s_2 can be computed following:

$$\begin{aligned} s_0 &= (I_0 + I_{\frac{\pi}{4}} + I_{\frac{\pi}{2}} + I_{\frac{3\pi}{4}})/2, \\ s_1 &= I_0 - I_{\frac{\pi}{2}}, \\ s_2 &= I_{\frac{\pi}{4}} - I_{\frac{3\pi}{4}}. \end{aligned} \tag{2}$$

Note that the integration of multiple light can be calculated as linear combination of their Stokes parameters. Then, ρ and ϕ are generated by Stokes elements as:

$$\rho = \frac{\sqrt{s_1^2 + s_2^2}}{s_0}, \phi = \frac{1}{2} \arctan \frac{s_2}{s_1}. \tag{3}$$

Also, s_1 and s_2 can be computed from s_0, ρ and ϕ by:

$$s_1 = s_0 \rho \cos(2\phi), s_2 = s_0 \rho \sin(2\phi). \tag{4}$$

4. Whitebox Attack on Glass Segmentation

Based on the novel locally controllable polarizing projection, we will show how to attack a polar-RGB based deep model, PGSNet for glass segmentation [26], in a whitebox manner. Assuming full access to the target model, our goal is to design an effective attack setting, find a stable perturbation, and project it onto the targeted scene in the physical world. The projected adversarial attacker will not be recognized by human eyes, since it appears to be uniformly white, yet can be captured by a polar-rgb camera. Under such manipulated inputs, the PGSNet model is induced to generate incorrect predictions.

4.1. Our Setting's Challenges

Fine-scaled perturbations like pixel-wise noises are extremely subtle and easy to be destroyed in a complicated physical world. Therefore, previous works apply large

perturbation patterns, like blocks [12], lines [11], triangles [38], for attacking image classification models. However, segmentation models predict pixel-wise classifications, which also apply advanced multiscale architectures, skip connections, and even self-attention modules, making them robust to adversarial attacks [3, 20]. Our polarization projection involves complex physical and optical properties, which makes the creation of adversarial samples highly complex and technical. Thus, we develop a perturbation pattern in the form of grids, to realize a more robust physical attack.

Normally, whitebox adversarial attacks in the physical world need to simulate comprehensive effects in the real world, including camera response function, camera noises, light decay, quantization effects, and so on [14, 19]. Therefore, in order to attack a polarization-based AI model, the most intuitive way of generating an effective perturbation in the whitebox manner is to simulate the complete transport of light. For us, polarization light travels from the polarizing projector to the object's surface, and finally to the polarization camera after being reflected by the object. However, it is not possible to acquire detailed scene geometry and accurate material parameters in the wild. So, we consider a simplified setup for whitebox attacks on glass segmentation. We can not only construct the most reliable simulation for polarization projection on glasses in the digital world but also generate adversarial examples of high effectiveness in the physical world.

4.2. Digital World Attack

Given a clean input $\mathbf{s}_b$ in the form of Stokes parameters and its binary label $y \in \mathbb{R}^{H\times W\times 1}$, where (H, W) denotes spatial resolution. PGSNet $f(\cdot)$ is trained to maximize the pixel-wise binary prediction accuracy, where $1/0$ represents the region that is/is not glass. The goal of our adversarial attack is to maximize the segmentation error with a projected adversarial perturbation, denoted as v, which is captured as $\mathbf{s} = \mathbf{s}(v)$. The problem is formulated as

$$\max_{v} \frac{1}{HW} \parallel y - f(g(\mathbf{s}_b + \mathbf{s}(v))) \parallel, \tag{5}$$

where $g(\cdot)$ denotes mapping from Stokes vectors to polarization cues by equation 3.

In general, adversarial attack algorithms generate adversarial examples by adding the gradient of error function w.r.t. $\mathbf{s}_b$, termed as $\nabla_{\mathbf{s}_b}\mathcal{L}$, and the perturbation is the division of clean image and its adversarial optimization result. However, the perturbation generated in this approach does not obey the physical property of our polarizing projection. For an optical adversarial system, we need to update perturbation directly following the gradient of the perturbation [14], i.e., the projection pattern v that will be fed into polarizing the projector, denoted as $\nabla_v\mathcal{L}$.

However, as mentioned, it is impossible to accurately construct a differentiable computation from v to $\mathbf{s}$, since the complicated polarization reflection. Thus, we try to generate perturbation by directly optimizing $\mathbf{s}$. We generate our adversarial example from a collection of real-world polarization images, termed as $\mathbf{S}_p = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_K\}$, captured directly by the targeted camera in the scene covered by the corresponding uniform projections $\mathbf{v}_p = \{v_1, v_2, ..., v_K\}$. For robust attacks in the real world, our projected perturbation is a map of grids, the value of each grid is assigned with a selected value, *e.g.*, v_i. With the relationship between the reflection $\mathbf{s}_i$ and the quantized projection v_i known, we can realize optimization using $\mathbf{S}_p$ rather than $\mathbf{v}_p$ to avoid complicated simulation of polarization reflection, indirect light effects, as well as the tone-mapping function of projector.

Further, we introduce a set of optimizable weights $\mathbf{\Omega} = \{\omega_1, \omega_2, ..., \omega_K\}$ on candidate images and use $SoftMax$ function to generate relative coefficients of each $\mathbf{s}_i$. Then we compute an adversarial example as:

$$\mathbf{s}_{ae} = \sum_{i}^{K} \frac{\exp(\omega_i/\tau)}{\sum_j^K \exp(\omega_j/\tau)}(\mathbf{s}_i - \mathbf{s}_b) + \mathbf{s}_b^*, \tag{6}$$

where τ is a temperature parameter to adjust the bias of relative weights. $\mathbf{s}_b^*$ denotes the augmented background image. Our optimization variable is the weights $\mathbf{\Omega}$. The problem in equation 5 is then reformulated as:

$$\max_{\mathbf{\Omega}} \frac{1}{HW} \parallel y - f(g(\mathbf{s}_{ae})) \parallel . \tag{7}$$

To deal with the problem in equation 7, we follow the negative gradient directions to update $\mathbf{\Omega}$ based on an iterative optimization approach:

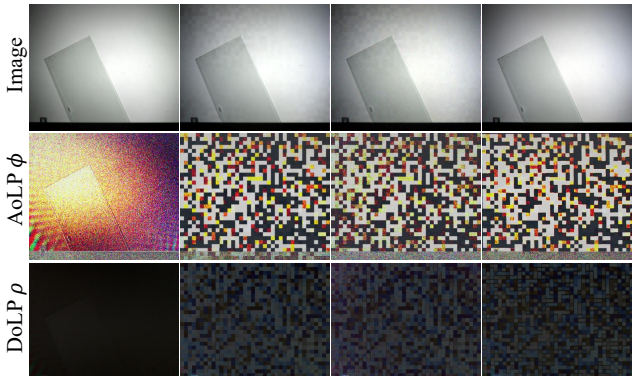
$$\mathbf{\Omega}^{t+1} \leftarrow \mathbf{\Omega}^t + \alpha\nabla_{\mathbf{\Omega}^t}\mathcal{L}(y, f(g(\mathbf{s}_{ae}))), \tag{8}$$

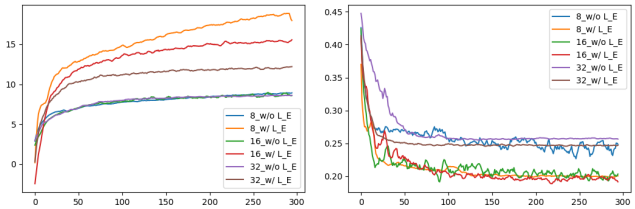
where α denotes the step size. After the optimization, we use $ArgMax$ to decide the final $\hat{\mathbf{\Omega}}$ and form an adversarial perturbation. In practical terms, to strike a balance between efficiency and effectiveness, we assembled a set of 17 candidate images. These images have source projection values that are uniformly discrete, ranging from 0 to 255. Further implementation details can be found in the supplementary material.

4.3. Adversarial Loss

With our adversarial example, we aim to maximize the error between predicted glass segmentation map $f(g(\mathbf{s}_{ae}))$ and label y, thus we first apply a Binary Cross Entropy loss. Moreover, we prefer to mislead the PGSNet to predict more non-glass pixels as positive, and vice versa, aligning with the approach outlined in [16]. With the prediction of adversarial example y_{ae}, $\mathcal{L}_E$ is termed as:

$$\mathcal{L}_E = \frac{1}{HW}\sum_{j\in y^n} y_{ae}^j - \frac{1}{HW}\sum_{j\in y^p} y_{ae}^j, \tag{9}$$





(a) $\mathcal{L}_{BCE}$ (b) IoU

Figure 5. Illustration of $\mathcal{L}_{BCE}$, and IoU during the optimization process.

where $y^{l\in\{n,p\}}$ denotes the set of pixels which are negative/positive (non-glass/glass) in the label y. y^n and y^p are disjoint, and the total number of pixels of $y^n \cup y^p$ is HW. Thus our final adversarial loss function is $\mathcal{L} = \mathcal{L}_{BCE} + \lambda\mathcal{L}_E$.

4.4. Augmentation for Attack in Real World

To generate more robust adversarial examples for the physical world attack, we follow the data augmentation strategy of EOT (Expectation Over Transformation) [5]. EOT employs a distribution of real-world degradations and transformations, enabling the generation of adversarial examples that are better suited for the complexities of the physical world. Given our specific focus on a scenario with a fixed, known camera and projector setup, transformations such as rotation and translation are not applicable in our case. We introduce Gaussian noise and apply Gaussian blur to simulate real-world degradation. Additionally, we employ a randomly sampled scale ratio to adjust the intensity of the background image $\mathbf{s}_b$, accounting for minor variations in environmental lighting conditions.

Table 1. Quantitative comparison using MAE and IoU in the digital and physical world. Random refers to perturbations that are randomly sampled, while Ours-k denote our optimized perturbation at a grid size of k.

	Digital world		Physical world	
	MAE↑	IoU ↓	MAE↑	IoU↓
Unpolarized	0.101	0.715	0.101	0.715
Random-8	0.204	0.461	0.167	0.509
Ours-8 w/o EOT	**0.745**	**0.185**	0.321	0.411
Ours-8 w/o $\mathcal{L}_E$	0.569	0.245	0.271	0.408
Ours-8	0.719	0.200	0.318	0.422
Random-16	0.246	0.489	0.287	0.449
Ours-16 w/o EOT	0.735	0.197	0.372	0.413
Ours-16 w/o $\mathcal{L}_E$	0.588	0.205	**0.404**	0.396
Ours-16	0.698	0.190	0.399	0.377
Random-32	0.329	0.440	0.257	0.474
Ours-32 w/o EOT	0.674	0.212	0.307	0.449
Ours-32 w/o $\mathcal{L}_E$	0.680	0.254	0.342	0.367
Ours-32	0.678	0.250	0.347	**0.361**

4.5. Experiments

To simplify our experiments, we use a specific setup with a co-located projector and camera. This configuration obviates the need for calibrating their relative poses, thereby facilitating an effortless alignment of the camera's view with the projection. We gather candidate images and background images in an indoor setting. In Figure 4, we show the visual comparison between the digital world simulation and real-world captures. The AoLP and DoLP images show that our simulation approach reconstructs realistic polarization reflections at extremely high precision. On the contrast, the modification of rgb images from adversarial perturbation is visually imperceptible, thus realize an undermined adversarial attack on polarization-based vision model.

Experiments were conducted across 11 scenes to validate the efficacy of the proposed method. At an image resolution of 612×512, we set the grid size for our perturbations to be 8, 16, and 32, respectively, that a smaller grid size yields a higher resolution for the perturbation. We apply MAE (Mean Absolute Error) and IoU (Intersection over Union) to characterize the prediction performance of the target model. For every scene, we run 300 iteration of optimizations, and $\lambda = 1$ for $\mathcal{L}_E$. The step size is set to $\alpha = 1$. We set τ as 0.3 and reduce it gradually for a simulation close to the candidate values. For a clear observation, the updates of $\mathcal{L}_{BCE}$ and IoU with different optimization settings are shown in Figure 5. The visual comparison shows the $\mathcal{L}_E$ effectively enforce erroneous predictions especially with high resolution perturbations (grid size 8).

Quantitative evaluations are summarized in Table 1. Based on the results, our polarizing perturbations can sig-

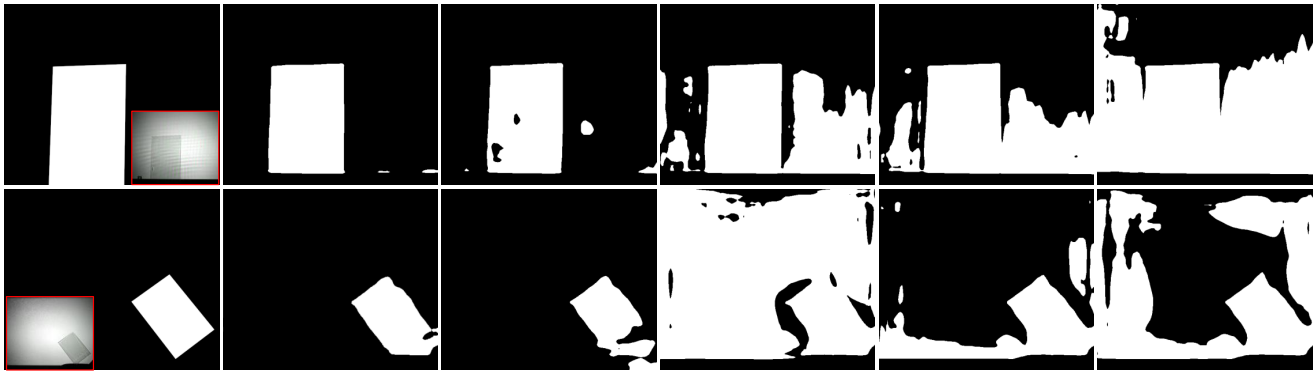


Figure 6. Visual comparison for adversarial attacks on the polarization-based glass segmentation model PGSNet [26]. This comparison ns, and polarized in the at a grid size of k. The

Figure 7. Representative ablation study regarding EOT and $\mathcal{L}_E$, using optimized perturbations at a grid size of 8.

nificantly weaken the accuracy of PGSNet. The comparison reveals that in the digital world, perturbations with higher resolution are more effective in misleading the targeted model, especially in scenarios without the application of EOT. Conversely, in the physical world, adversarial perturbations with a lower resolution (grid size 32) exhibit superior performance. We attribute the improvement to the inherent robustness provided by grid-based perturbation. This robustness ensures effective transferability, even in the face of minor degradations. Furthermore, the use of (EOT) prevents overfitting to the input data and enhances the transferability of adversarial examples to real-world scenarios. Notably, our optimized adversarial examples outperform randomly generated perturbations with a great margin, and the introduction of our proposed $\mathcal{L}_E$ loss further amplifies attack efficacy in both digital and physical realms.

We illustrate results of adversarial examples in two grid sizes, 8 and 32, as shown in Figure 6. When compared with predictions derived from inputs illuminated by unpolarized projection, the results highlight the efficacy of our polarizing projection in undermining the performance of PGSNet in both the digital and physical worlds. Notably, even perturbations that are randomly sampled can degenerate the model's performance. Furthermore, perturbations synthesized via our optimization technique consistently outperform random perturbations. Particularly in physical world attacks, our method benefits substantially from the integration of Expectation Over Transformation (EOT) and the $\mathcal{L}_E$ loss function, resulting in robust and pronounced attack performance. Although minimal visual textures are discernible to the human eye, the polarization properties undergo significant alterations due to our perturbation projections. This approach effectively achieves both stealthiness and attack efficacy. Additional experimental details are provided in our supplementary material.

Visual comparisons of the physical world attack for the ablation study are also shown in Figure 7, which indicate the significance of the proposed technique. The applied EOT enhances the robustness of the projection perturbation effectively and proposed $\mathcal{L}_E$ further boosts the attack performance.

5. Whitebox Attack on Deep SfP

We have expanded our locally controllable polarizing projection technique to another key area of polarization imaging: shape estimation. Polarization imaging is inherently adept at capturing cues related to object geometries. In line with this, SfP-wild [23] suggests leveraging deep priors from a large-scale polarization image dataset to estimate normal maps from single-shot images taken in the wild. Compared to RGB-only-based normal estimation, Lei's model derives significant advantages from polarization cues, enabling it to discern false geometries, such as scenes printed in a photograph. Importantly, with deep priors derived from extensive datasets captured in the wild, the model avoids the ambiguity issue [13] and circumvents

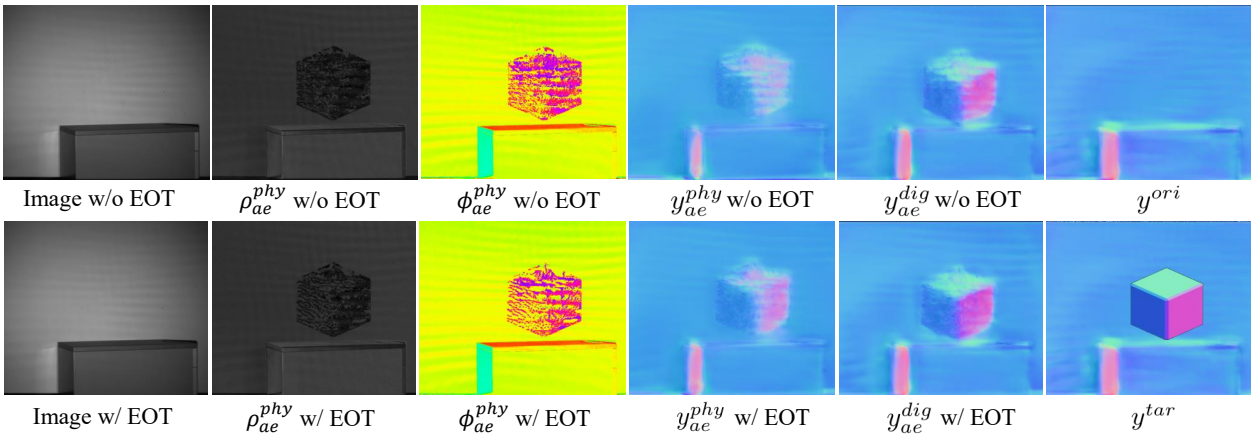


Figure 8. Visual comparison for adversarial attacks on the shape esimation model SfP-wild [23].

lighting constraints [10, 18]. However, our experiments demonstrate instability when relying on polarization. Our polarizing projection can mislead the state-of-the-art deep Shape-from-Polarization model. This allows the fake creation of arbitrary shapes through perturbations that are imperceptible to the human eye.

Our objective is to project a polarizing perturbation onto the background, aiming to deceive the model into estimating shapes of non-existent 'objects'. Starting with an original estimation under a uniformly linear polarized projection, we superimpose the background normal map with a hypothetical object, such as a cube, to serve as our target for the attack. We follow the settings described in Section 4 to optimize the perturbation pattern within the target region. For the optimization process, we simply employ the MAE loss [23] and update the perturbation by gradient descent. We introduce a high-resolution perturbation with a grid size of 2 and also incorporate the EOT methodology [5] to ensure an effective adversarial perturbation in real world attacks.

In Figure 8, we present the adversarial examples resulting from the physical world attack, alongside the network outputs $y_{ae}^{l\in\{phy,dig\}}$ corresponding to physical and digital world attacks. Here, y^{ori} represents the estimation obtained with the background illuminated by linear polarized projection, while y^{tar} denotes the label of our targeted adversarial attack. As indicated by the intensity, ρ_{ae}^{phy} and ϕ_{ae}^{phy} of adversarial examples, our perturbation focuses on modifying the polarization within the target region. In the digital domain, even a limited polarizing reflection proves adequate for generating a detailed normal map. While certain inherent challenges, such as noise and quantized signals, manifest in the real world, the attack leveraging EOT still yields results that closely align with the simulations. Please refer to our supplementary material for more evaluation.

In addition to the aforementioned experiments, we also conducted tests targeting the polarization-based color constancy algorithm [28] and human pose and shape estimation model [39]. Further details on these experiments are provided in our supplementary material.

6. Research Ethics and Limitations

This study originates from our curiosity on the potential vulnerability of polarization-based vision algorithms in the digital space. In line with existing researches on adversarial attacks, this study is intended to offer a timely warning on the potential vulnerability of polarization-based AI.

The most obvious limitation we found lies in the relative low luminance of the projector, and the attack success rate will be low in bright environment. However, it is highly effective in indoor or low-light outdoor scenarios. Further protection measures, such as adversarial training, data enhancement, or introducing activate illuminations should be considered.

7. Conclusion

Polarization has been utilized for a variety of computer vision tasks. We have shown that, similar to the well-known vulnerability of RGB-based vision, the performance of polarization-based vision algorithms, such as glass segmentation and shape estimation, can be manipulated, maybe in a potentially harmful way. Our adversarial attackers are physically realized by using an adapted one-chip LCD projector, which allows locally controllable polarizing projection. Our method is visually friendly, thus poses realistic concerns on the reliability of polarization-based AI. We hope this study will arouse attentions on the potential risks of polarization-based vision.

Acknowledgement This research was supported in part by JSPS KAKENHI Grant Numbers 22H00529, 20H05951, 23H03420, JST-Mirai Program JPMJMI23G1, and ROIS NII Open Collaborative Research 2023-23S1201.

References

[1] Naveed Akhtar and Navid Kardan. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 1

[2] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021. 1

[3] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018. 5

[4] Mark Ashdown, Takahiro Okabe, Imari Sato, and Yoichi Sato. Robust content-dependent photometric projector compensation. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 6–6. IEEE, 2006. 3

[5] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 2, 6, 8

[6] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 554–571. Springer, 2020. 3

[7] Seung-Hwan Baek, Daniel S Jeon, Xin Tong, and Min H Kim. Simultaneous acquisition of polarimetric svbrdf and normals. *ACM Trans. Graph.*, 37(6):268–1, 2018. 3

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[9] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1558–1567, 2017. 3

[10] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15567–15576, 2021. 1, 3, 8

[11] Ranjie Duan, Xiaofeng Mao, A Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun Yang. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16062–16071, 2021. 2, 5

[12] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 5

[13] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Polarimetric normal stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 682–690, 2021. 1, 3, 7

[14] Abhiram Gnanasambandam, Alex M Sherman, and Stanley H Chan. Optical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 92–101, 2021. 2, 3, 5

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[16] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 308–325. Springer, 2022. 5

[17] Chengyin Hu and Weiwen Shi. Adversarial color projection: A projector-based physical attack to dnns. *arXiv preprint arXiv:2209.09652*, 2022. 3

[18] Inseung Hwang, Daniel S Jeon, Adolfo Munoz, Diego Gutierrez, Xin Tong, and Min H Kim. Sparse ellipsometry: portable acquisition of polarimetric svbrdf and shape with unstructured flash photography. *ACM Transactions on Graphics (TOG)*, 41(4):1–14, 2022. 3, 8

[19] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 962–969, 2019. 2, 5

[20] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8828–8838, 2020. 5

[21] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 2

[22] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1750–1758, 2020. 1, 3

[23] 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. 1, 2, 3, 7, 8

[24] Mengke Li, Naifu Yao, Sha Liu, Shouqing Li, Yongqiang Zhao, and Seong G Kong. Multisensor image fusion for automated detection of defects in printed circuit boards. *IEEE Sensors Journal*, 21(20):23390–23399, 2021. 1

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[26] 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, 4, 7

[27] Shao Mingqi, Xia Chongkun, Yang Zhendong, Huang Junnan, and Wang Xueqian. Transparent shape from single polarization images. *arXiv preprint arXiv:2204.06331*, 2022. 3

[28] 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. 1, 3, 8

[29] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 2

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[31] Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, and Yasushi Yagi. Recovering inner slices of translucent objects by multi-frequency illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5464–5472, 2015. 3

[32] Silvia Tozza, Dizhong Zhu, William AP Smith, Ravi Ramamoorthi, and Edwin R Hancock. Uncalibrated, two source photo-polarimetric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5747–5760, 2021. 3

[33] Hui Wei, Tang Hao, Xuemei Jia, Hanxun Yu, Zhubo Li, Zhixiang Wang, Shin'ichi Satoh, and Zheng Wang. Physical adversarial attack meets computer vision: A decade survey. *arXiv preprint arXiv:2209.15179*, 2022. 2

[34] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 1–17. Springer, 2020. 2

[35] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 665–681. Springer, 2020. 2

[36] Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, and Ping Tan. Polarimetric dense monocular slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3857–3866, 2018. 3

[37] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[38] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15345–15354, 2022. 2, 5

[39] 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. 8