Understanding of the SPAA paper

SPAA: Stealthy Projector-based Adversarial Attacks on Deep Image

(Dated: April 15, 2023)

I. BACKGROUND

Three common ideas of adversarial attacks:

Digital attack, which altering input digital image by adding small perturbations.

Physical attack, which placing manufactured adversarial objects.

Projector-based attack, which is different from the two kinds above and has following attentions.

(a) Using SAR technique to implement it.

(b) Based on optics, no need to place adversarial objects,

(c) Can be performed transiently and dynamically.

(d) There are two common methods about projector-based attack. One is first proposed by Nichols and Jasper in paper Projecting trouble: Light based adversarial attacks on deep learning classifiers. The other is dividing into two steps: performing digital attacks and then using projector compensation techniques to implement it. However, Both of them have obvious drawbacks. The fronter use DE techniques to implement it. But when it comes to high resolution case, it will be hard to run in parallel and a efficiency bottleneck occurs. The later use the method of digital attacks at first but digital attacks may generate physically implausible adversarial examples that cannot be produced by a projector.

(e) To implement the stealthiness, the L$_p$ is commonly used, but the paper Towards large yet imperceptible adversarial image perturbations with perceptual color distance shows perceptual color distance maybe better.

II. METHODS

The image classifier

$$\underset{i}{\operatorname{argmax}}\, f_i(I+\delta) \begin{cases} = t & \text{targeted} \\ \neq t_{\text{true}} & \text{untargeted} \end{cases}$$
$$\text{subject to} \quad \mathcal{D}(I, I+\delta) < \epsilon,$$

Projector-based attack

$$\underset{i}{\operatorname{argmax}}\, f_i(\pi_c(l+\delta_l, s)) \begin{cases} = t, & \text{targeted} \\ \neq t_{\text{true}} & \text{untargeted} \end{cases}$$
$$\text{subject to} \quad \mathcal{D}\left(\pi_c(l+\delta_l, s), \pi_c(l, s)\right) < \epsilon$$

Projector’ s projection function

$$\underset{i}{\operatorname{argmax}}\, f_i\left(I_{x'} = \pi(x', l, s)\right) \begin{cases} = t, & \text{targeted} \\ \neq t_{\text{true}} & \text{untargeted} \end{cases}$$
$$\text{subject to} \quad \mathcal{D}\left(I_{x'}, I_{x_0}\right) < \epsilon,$$

CompenNet++

An intuitive solution to digitally attack the camera-captured scene image under normal light first, then use a projector compensation method to find its corresponding projector input image. However, it cannot address occlusions and those occlusive regions may become blurry after compensation.

PCNet

(a) Only the projector input image x is varied.

$$\hat{I}_x = \hat{\pi}(x, I_s, I_{\mathrm{m}})$$

(b) Using a plain gray image to provide some illumination.

(c) Using a projector direct light mask to exclude occluded pixels.

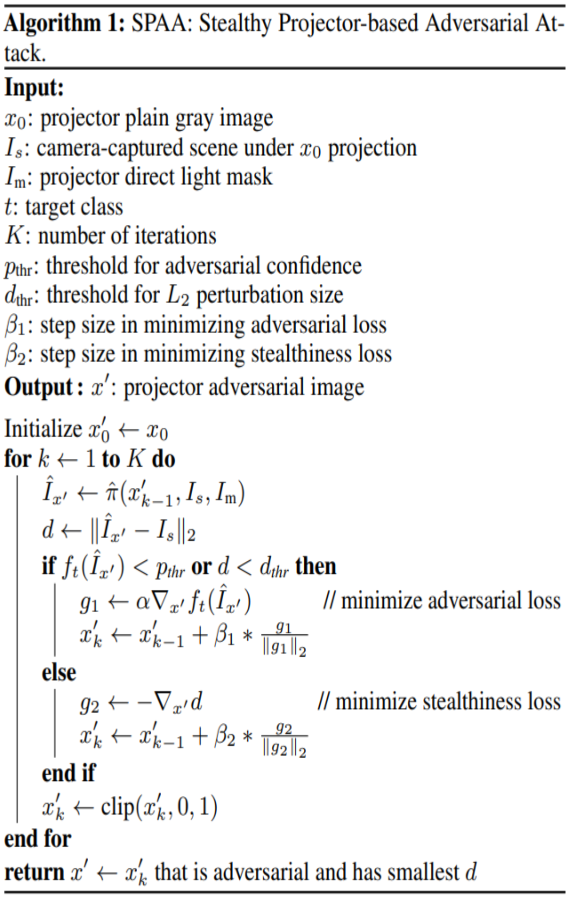
(d) Two subnet: Warping Net and Shading Net. Warping Net consists of a learnable affine matrix , TPS parameters and a grid refinement network. Shading Net consists of a two-branch encoder-decoder structure: middle encoder , backbone encoder.

(e) PCNet parameters can be trained using image reconstruction loss L.

$$\theta = \underset{\theta'}{\operatorname{argmin}} \sum_i \mathcal{L}\left(\hat{I}_{x_i} = \hat{\pi}_{\theta'}(x_i, I_s, I_{\mathrm{m}}),\ I_{x_i}\right)$$

SPAA

Using two thresholds to evaluate current image. one threshold is for adversarial confidence, another threshold is for L2 perturbation size.



Algorithm 1: SPAA: Stealthy Projector-based Adversarial Attack.

```
Input:
$x_0$: projector plain gray image
$I_s$: camera-captured scene under $x_0$ projection
$I_{\mathrm{m}}$: projector direct light mask
$t$: target class
$K$: number of iterations
$p_{\mathrm{thr}}$: threshold for adversarial confidence
$d_{\mathrm{thr}}$: threshold for $L_2$ perturbation size
$\beta_1$: step size in minimizing adversarial loss
$\beta_2$: step size in minimizing stealthiness loss
Output: $x'$: projector adversarial image

Initialize $x'_0 \leftarrow x_0$
for $k \leftarrow 1$ to $K$ do
    $\hat{I}_{x'} \leftarrow \hat{\pi}(x'_{k-1}, I_s, I_{\mathrm{m}})$
    $d \leftarrow \|\hat{I}_{x'} - I_s\|_2$
    if $f_t(\hat{I}_{x'}) < p_{thr}$ or $d < d_{thr}$ then
        $g_1 \leftarrow \alpha \nabla_{x'} f_t(\hat{I}_{x'})$        // minimize adversarial loss
        $x'_k \leftarrow x'_{k-1} + \beta_1 * \frac{g_1}{\|g_1\|_2}$
    else
        $g_2 \leftarrow -\nabla_{x'} d$        // minimize stealthiness loss
        $x'_k \leftarrow x'_{k-1} + \beta_2 * \frac{g_2}{\|g_2\|_2}$
    end if
    $x'_k \leftarrow \mathrm{clip}(x'_k, 0, 1)$
end for
return $x' \leftarrow x'_k$ that is adversarial and has smallest $d$
```

Stealthiness

(a) Inspired by digital attack algorithms PerC-AL and DDN, but they are digital attacks, which is different from projector-based adversarial attack.

(b) Try to improve the algorithm by optimizing the weighted sum of adversarial and stealthiness losses. However, it doesn't work.

III. EXPERIMENTS

Evaluate settings

(a) Evaluate stealthy projector-based attack methods by targeted and untargeted attack success rates.

(b) Evaluate stealthiness measured by similarities between the camera-captured scene and the camera-captured scene under adversarial projection using L_2, L_∞, E, SSIM.

(c) Compared with One-pixel DE and PerC-AL+CompenNet++

(d) Better efficiency than DE. Beacuse DE involves a non-parallelizable real project-and-capture process.

Evaluate results

(a) SPAA and PCNets have higher attack success rates and stealthiness. One-pixel DE has low targeted attack success rates. It only perturbs a 41 Œ 41 projector image block and it makes camera-captured images have strong square patterns. PerC-AL+CompenNet++ produces a blurry bucket-like projection pattern and may produce strong adversarial patterns in the bucket shadow where the projector is unable to project to the occluded region.

Perturbation size threshold When the perturbation size threshold is higher, the attack success rates will be higher, but perturbation sizes will be also higher, which means lower stealthiness.

PCNet components Try the settings of different norms and without direct light mask and find that direct light mask actually makes algorithms better.

IV. THOUGHTS

Different pose

(a) Projector-based attack can be performed transiently and dynamically.

(b) Because PCNet can be trained offline, it requires only one online project-and-capture process for stealthy projector-based attacks.

(c) Maybe some tracking techniques.(still need improve)