利用者名：コンピュータサイエンス学部　准教授　宇田 隆哉

Scopus

Title: Invisible Cloak to AI Recognition from All Horizontal Directions by Adversarial Patch
（アドバサリアルパッチにより水平方向すべての向きに対して AI の認識から透明になれるマント）
Authors: Takumi Imaeda and Ryuya Uda
（今枝 拓未（東京工科大学大学院）、宇田 隆哉（東京工科大学 准教授））
Journal: Proc. 18th International Conference on Ubiquitous Information Management and Communication, 2024
掲載年月：2024 年 2 月

研究概要：アフィン変換と TPS マッピングによる処理を施した Adversarial Patch を、マントに 5 枚貼り付けることで、このマントを着用した人間が水平方向 360 度のどの角度から撮影されても、人工知能の目を欺き、人間がそこにはいないと誤認識させられます。

研究背景：Adversarial Patch というものがあります。このパッチ当てたものを人工知能に見せると、その人工知能はそのものを誤認識してしまうというものです。Brown らが現実世界に適用可能な Adversarial Patch を提案しました。Thys らは Adversarial Patch を物理的なボードに実装し、そのボードを持っている人は、AI から人と認識されないようにしました。Xu らは Adversarial Patch を T シャツにプリントして TPS マッピングにより歪み耐性を持たせました。しかし、いずれの研究も、人間が正面から撮影されることを前提としていまして、斜めの方向から撮影されるとパッチの効果がうまく発揮されません。そこで、我々の研究室から、金井らの透明マントが発表されます。パッチにアフィン変換を行って撮影角度耐性を持たせ、TPS マッピングにより歪み耐性も持たせました。ただし、金井らのパッチはマントの前面に 3 枚あるのみですので、正面から見て少し斜め方向から撮影される場合には効果がありますが、真横や真後ろから撮影される場合には効果がありません。このような経緯を経て、水平方向すべての向きからの撮影に対して、効果があるマントを作ることにしました。

研究成果：マントの周囲に 5 枚のパッチを均等に貼ります。この 5 枚のパッチは、それぞれが均等に撮影角度耐性と歪み耐性を持っていまして、水平方向すべての向きに対してほぼ同等の妨害効果を発揮できるようになっています。このマントを、情報処理学会のコンピュータセキュリティ（CSEC）研究会で発表したのですが、そこで問題が発生しました。我々のパッチは、Thys らのものを元にしています。論文の図 4 が Thys らのパッチなのですが、いくつかのパラソルの下に何人かの人間がいるように見えなくもありません。そして、このパッチを人物の写真の上に貼り付けてみたものが論文の図 5 なのですが、パッチのところに赤い枠線が表示されてます。人工知能による人間の認識には YOLO (You Only Look Once)を使っているのですが、YOLO では人間を識別した箇所に赤い枠線を表示します。つまり、パッチを当てた人は、パッチの効果により人間とは認識されていないのですが、YOLO がパッチの模様の中に人間を見つけてしまい、そこに人間がいることにされてしまっているのです。Thys らのパッチでは、パラソルの下に人間がいるように見えるというのは、人間による主観だけでなく、人工知能にとってもそう見えるということです。CSEC の論文では、このパッチでマントを作ってしまいましたので、稀に人間がいると判定されてしまいました。Thys らのパッチは与えられた初期画像から作り出されるのですが、この初期画像をリンゴに変えたところ、論文の図 10 のようなパッチになりました。これで、パッチの中に人間のようなものは

Fig. 5. Example of human detection in a patch image.

Fig. 4. Adversarial Patch by Thys et al.

Fig. 10. Adversarial patch with initial image of apples.

なくなりましたので、妨害効果が完全になると思い、実験をしてみました。結果が論文の図 8 と図 9 です。

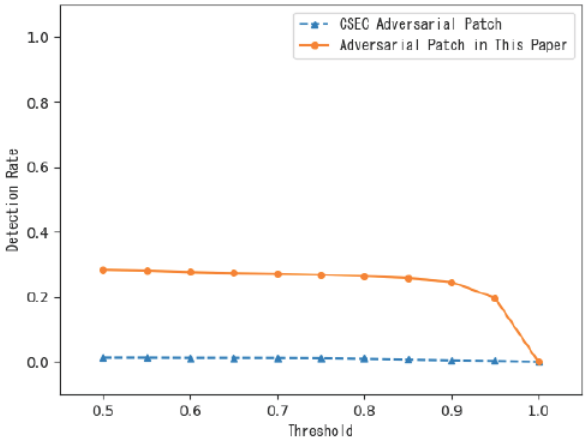

Fig. 8. Human detection rate from all angles.

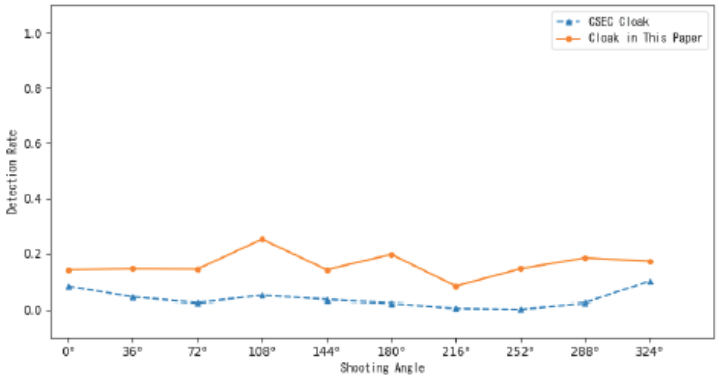

Fig. 9. Human detection rate from each angle with wrinkles.

社会的・学術的なポイント：確かに、パッチの中に人間のようなものは認識されなくなりましたが、パッチそのものの効果が落ちてしまい、図 8 でも図 9 でも、CSEC の論文のマントよりも妨害効果が低くなってしまいました。パッチの効果は初期画像に依存するようで、いくつかの初期画像で実験してみたところ、シアン一色のものから作ったパッチが一番効果が高いとなりました。この論文での研究はここまでですが、実は研究が進んでマントが完成しています。そちらの論文にご期待ください。

用語解説：

TPS マッピング：Thin Plate Spline マッピングの略で、布を歪ませたときに、歪んでいない布の特定の位置にある色から、歪んだ場合の特定の位置にある色を補間するマッピング処理に使用しています。

アフィン変換：画像を平行移動させたり拡大・縮小させたり回転させたりなどする変換のこと。