最適な遠隔敵対的パッチの形状は四角ではない

大西 健斗[1,a)]　中井 綱人[1]

概要：本稿では，畳み込みニューラルネットワーク (CNN) を利用した物体検知において，誤認識を引き起こす効果の高い遠隔敵対的 patch の形状の理論的考察を行う．CNN に対する強大な脅威の一つとして，敵対的な特徴を持つ patch を利用した攻撃方法がある．しかし，既存研究では，長方形の敵対的 patch による攻撃が主であり，その形状についてはあまり考察されてこなかった．本稿では，Oonishi らが提案した敵対的 patch 効果の拡散モデルに基づき，最適な遠隔敵対的 patch の形状について理論的解析を行う．まず，敵対的 patch の効果がほぼ同心円状に拡散することを示すとともに，画像の中心に強い影響を与える遠隔敵対的 patch は円の一部となることを理論的に示す．さらに，最適な遠隔敵対的 patch を配置するアルゴリズムを提案し，得られる形状が理論的解析と同じく円の一部であることを示すとともに，平均検出率が，長方形の場合の 12.81%から 9.27%まで低下することを示す．最後に，遠隔敵対的 patch の動的な生成を行った場合にも，遠隔敵対的 patch が円の一部となることを，実験により示す．

キーワード：CNN，YOLO，Adversarial patch，理論的解析

Optimal Remote Adversarial Patches are NOT Rectangles

Kento Oonishi[1,a)]　Tsunato Nakai[1]

Abstract: This paper gives an analysis of strong remote adversarial patches on Convolutional Neural Network (CNN)'s object detection. Adversarial patches are one of severe threats for CNNs. However, researchers have mainly considered rectangle adversarial patches, and an optimal shape has not been fully considered. This paper gives an optimal shape of remote adversarial patches based on Oonishi's diffusion model. First, we show that an effect of adversarial patches spreads almost concentrically. We then show that remote adversarial patches with a part of a circle give strong effects on the center of images. We also propose an algorithm for calculating optimal remote adversarial patches. Our algorithm's output is a part of a circle, which is the same as our theoretical analysis. Moreover, an average detection rate is 9.27%, which is lower than 12.81% by previous rectangle adversarial patches. Finally, we show that adversarial patches generated dynamically is a part of a circle.

Keywords: CNN，YOLO，Adversarial patch，Theoretical analysis

1. はじめに

1.1 研究背景

現在，人工知能による自動化が，様々な分野で進行中である．特に，物体検知技術は [7] は，人工知能が次処理の判断を行う上で必須の技術である．ここで，状況を正確に捉えた処理を行うには，正確な物体検知を行うことが極めて重要である．以上の背景を踏まえ，近年，物体検知の正確性を高める研究が進められてきた．

しかし，上記の物体検知には，敵対的 patch 攻撃 [1], [3], [4], [5], [6], [8], [9], [10], [12] の脅威がある．敵対的 patch 攻撃は，入力画角上に特殊な画像 (敵対的 patch) を含めることで，物体検知システムを阻害する攻撃方法である．この敵対的 patch 攻撃は，画角内に patch を配置すれば成立する，という点で，物体検知システムに対する現実的な脅威である．したがって，物体検知システムが敵対的 patch

[1] 三菱電機株式会社情報技術総合研究所
Mitsubishi Electric
[a)] Onishi.Kento@ap.MitsubishiElectric.co.jp

攻撃に対してどの程度脆弱であるかを見積もるのは極めて重要である．

敵対的 patch 攻撃の脅威を現実的な条件下で見積もるにあたり，事前に決められた形状で敵対的 patch を配置する攻撃方法が多数提案されている [3], [5], [6], [8], [9], [10]．現在までに提案された敵対的 patch の多くは，対象物上に敵対的 patch を配置することで，誤認識を引き起こす [3], [9], [10]．これらの攻撃方法は，対象物自身が物体検知を回避する目的で利用される．対象物上に敵対的 patch を配置する攻撃方法とは異なり，対象物外に，遠隔で敵対的 patch を配置する攻撃方法も多数提案されている [5], [6], [8]．これらの攻撃方法は，物体検知を回避する目的を持たない対象物に対しても攻撃可能であるため，より強力な脅威となりうる．したがって，本稿では，遠隔で敵対的 patch を配置する攻撃方法について考察を行う．

遠隔で敵対的 patch を配置する攻撃方法への理論的なアプローチとして，Oonishi ら [6] は，敵対的 patch の効果の拡散モデルを利用した考察を行った．Oonishi らは，3×3 のフィルターサイズを持つ畳み込み層によって，敵対的 patch の効果が拡散する，というモデルの下で，敵対的 patch の影響を評価した．その上で，画像内に複数の敵対的 patch を配置すると，敵対的 patch の効果が画像全体に拡散し，敵対的 patch 攻撃の効果が高まることを示した．Oonishi らは，様々な敵対的 patch の配置方法について考察を行ったが，長方形の敵対的 patch を恣意的に配置しており，その配置方法，特に，敵対的 patch の形状には改良の余地がある．したがって，敵対的 patch の効果を高める形状についての考察は極めて重要である．

1.2 研究成果

本稿では，YOLOv2 [7] に対し，Oonishi らの既存研究 [6] に基づき，より強力な敵対的 patch の生成方法の提案を行う．本稿では，Oonishi らの既存研究と同じく，敵対的 patch の総面積が単一遠隔 patch とほぼ同一となる条件下で，最適な遠隔敵対的 patch の配置法を提案する．本稿の成果は，以下の通りである．特に，画像の中心部分にある物体に対する攻撃を想定した理論的解析を行う．

- **研究成果 1:** 敵対的 patch 効果の拡散の近似的な理論的解析 (3 節)
- **研究成果 2:** 最適な遠隔敵対的 patch の配置アルゴリズムの提案 (4 節)
- **研究成果 3:** 動的な遠隔敵対的 patch の生成による理論的解析の実験的検証 (5 節)

本稿では，まず，3 節において，敵対的 patch 効果の拡散の近似的な理論的解析を行う．はじめに，1 つの pixel のみに敵対的 patch を配置した場合の拡散効果について定量的な評価を行う．本稿では，厳密な評価を行った後，解析を容易にするために，近似的な評価を行う．上記の理論的解析の結果，敵対的 patch の効果は，ほぼ同心円状に拡散し，中心となる pixel から離れるごとに減衰することを示す．上記は，Yu ら [11] によって実験的に示されているが，本稿では，理論的な根拠を示している．なお，上記の「ほぼ」は，中心となる pixel に近い範囲で成立する，という意味である．さらに，YOLOv2 において，画像中心部分の領域に高い効果を与える pixel について理論的考察を行う．上記の解析により，画像の中心点に近い pixel ほど，画像中心部分の領域に高い効果を与えることが可能であることを示す．

次に，4 節において，最適な遠隔敵対的 patch の配置アルゴリズムの提案を行う．本稿では，Oonishi らの拡散モデルに基づいて，各 pixel が画像中心部分の領域における影響値を計算する．その上で，画像中心部分の領域に置ける影響値が大きい pixel から順に pixel を選択し，最適な遠隔敵対的 patch を与える領域を計算するアルゴリズムを提案する．以上の提案アルゴリズムを利用した計算の結果，最適な遠隔敵対的 patch は，ほぼ，円の一部となることを示す．さらに，平均検出率が，長方形の場合の 12.81%から，9.27%まで低下することを示す．

最後に，5 節において，遠隔敵対的 patch の動的な生成を行い，遠隔敵対的 patch が，ほぼ，円の一部となることを実験により示す．5 節は，3，4 節とは異なり，Oonishi らの敵対的 patch 効果の拡散モデルには基づいていない．5 節は，Huang ら [4] や Zhu ら [12] の手法と同様，事前に敵対的 patch の位置を固定せず，最適な遠隔敵対的 patch の形状を学習する．本稿では，以上の実験において，最適な遠隔敵対的 patch の形状に関する学習を行ったとしても，3，4 節と同様，円の一部の形状を持つ敵対的 patch が生成されることを示すことで，Oonishi らの拡散モデルに基づいた理論的解析の正当性を示す．

2. 準備

本節では，まず，2.1 節で，本稿で導入する記法について説明する．その後，2.2 節で，本稿の攻撃対象である YOLOv2 [7] について説明する．最後に，2.3 節で，Oonishi らの敵対的 patch 効果の拡散モデル [6] を説明する．

2.1 本稿で用いる記法

本稿における，log の底は e とする．本稿では，Saha らの既存研究 [8] と同じく，公開データセットである PASCAL VOC データセット [2] を入力画像として用いる．本稿では，画像サイズは 416×416 とし，画像の横軸方向を x 成分，縦軸方向を y 成分とする．その上で，左上の pixel を $(0, 0)$ とし，左から右に移動するごとに x 座標の値が増加，上から下に移動するごとに y 座標の値が増加するものとする．以上より，画像の各 pixel は，$\{(x, y) \,|0 \leq x \leq 415, 0 \leq y \leq 415\}$ で表される．

表 1 VOC データセット用に作成された YOLOv2 の CNN
Table 1 YOLOv2's CNN for the VOC dataset

Layer	フィルター数	フィルターのサイズ	ストライド	入力サイズ	出力サイズ
0 conv	32	3 × 3	1	416 × 416 × 3	416 × 416 × 32
1 max	–	2 × 2	2	416 × 416 × 32	208 × 208 × 32
2 conv	64	3 × 3	1	208 × 208 × 32	208 × 208 × 64
3 max	–	2 × 2	2	208 × 208 × 64	104 × 104 × 64
4 conv	128	3 × 3	1	104 × 104 × 64	104 × 104 × 128
5 conv	64	1 × 1	1	104 × 104 × 128	104 × 104 × 64
6 conv	128	3 × 3	1	104 × 104 × 64	104 × 104 × 128
7 max	–	2 × 2	2	104 × 104 × 128	52 × 52 × 128
8 conv	256	3 × 3	1	52 × 52 × 128	52 × 52 × 256
9 conv	128	1 × 1	1	52 × 52 × 256	52 × 52 × 128
10 conv	256	3 × 3	1	52 × 52 × 128	52 × 52 × 256
11 max	–	2 × 2	2	52 × 52 × 256	26 × 26 × 256
12 conv	512	3 × 3	1	26 × 26 × 256	26 × 26 × 512
13 conv	256	1 × 1	1	26 × 26 × 512	26 × 26 × 256
14 conv	512	3 × 3	1	26 × 26 × 256	26 × 26 × 512
15 conv	256	1 × 1	1	26 × 26 × 512	26 × 26 × 256
16 conv	512	3 × 3	1	26 × 26 × 256	26 × 26 × 512
17 max	–	2 × 2	2	26 × 26 × 512	13 × 13 × 512
18 conv	1024	3 × 3	1	13 × 13 × 512	13 × 13 × 1024
19 conv	512	1 × 1	1	13 × 13 × 1024	13 × 13 × 512
20 conv	1024	3 × 3	1	13 × 13 × 512	13 × 13 × 1024
21 conv	512	1 × 1	1	13 × 13 × 1024	13 × 13 × 512
22 conv	1024	3 × 3	1	13 × 13 × 512	13 × 13 × 1024
23 conv	1024	3 × 3	1	13 × 13 × 1024	13 × 13 × 1024
24 conv	1024	3 × 3	1	13 × 13 × 1024	13 × 13 × 1024
25 route 16					
26 conv	64	1 × 1	1	26 × 26 × 512	26 × 26 × 64
27 reorg	–	–	2	26 × 26 × 64	13 × 13 × 256
28 route 27 24					
29 conv	1024	3 × 3	1	13 × 13 × 1280	13 × 13 × 1024
30 conv	125	1 × 1	1	13 × 13 × 1024	13 × 13 × 125
31 detection					

2.2 YOLOv2 [7]

YOLOv2 は，一度の CNN 処理で物体検知が可能なネットワークである．本稿では，416×416 のサイズの画像を入力とする，VOC データセット [2] 用に作成された YOLOv2 を扱う．YOLOv2 の CNN は，表 1 の通りである．

表 1 の CNN は，13×13 のセルごとに，125 個の値を出力する．この 125 個の値は，それぞれ 25 個の値の組を持つ 5 つの bounding box のデータで構成されている．それぞれの anchor box に含まれる 25 個の値は，

- bounding box の位置 (4 値)
- 対象物が存在する確率 (1 値)
- 各対象物の存在確率 (20 値，以降，クラス値と呼ぶ)

で構成され，対象物が存在する確率とクラス値の積が検出率となる．

2.3 Oonishi らの敵対的 patch 効果の拡散モデル [6]

Oonishi らは，敵対的 patch の拡散モデルを生成し，敵対的 patch の効果を影響値として計算することで，敵対的 patch の効果を評価した．影響値は，YOLOv2 の出力である 13×13 サイズの画像の各 pixel で計算され，この影響値が大きいほど patch の効果が高いことを示す．具体的な計算方法は，以下の通りである．

- 影響値の初期値について，敵対的 patch 部分は 1，それ以外の部分は 0 とする．
- フィルターサイズが 3×3 の畳み込みで，敵対的 patch の影響値を，周囲 9 マスの影響値の平均値へ更新する．
- 敵対的 patch の影響値の更新に必要な 9 マスに画像の外側が含まれる場合，それらの部分の影響値は 0 として計算する．

なお，以上の拡散モデルでは，3×3 の畳み込み層以外に関する操作を省略しており，本稿では，これらの操作を以下の通りに補完する．

- フィルターサイズが 1×1 の畳み込み層では，更新を行わない．
- フィルターサイズが 2×2 のプーリング層では，計算に利用する 4 マスの平均値を計算する．
- 第 25 層の “route 16” では，第 16 層の出力を，第 25 層の出力とする．
- 第 28 層の “route 27 24” では，第 27 層の出力と第 24 層の出力の合計値を，第 28 層の出力とする．

本稿では，以上の拡散モデルの下で，敵対的 patch 効果の理論的解析を行うとともに，より強力な敵対的 patch の配置方法について議論を行う．

3. 研究成果 1: 敵対的 patch 攻撃の影響についての理論的解析

本節では，敵対的 patch の拡散モデルに基づき，敵対的 patch の画像全体への影響の理論的解析を行う．はじめに，3.1 節で，1 pixel の敵対的 patch の影響がどのように拡散するかを近似的に計算する．上記の近似的な計算により，敵対的 patch の影響は，元々の 1 pixel を中心として，ほぼ同心円状に拡散することを示すとともに，中心となる pixel から遠くなるほど弱まることを示す．さらに，3.2 節で，YOLOv2 での敵対的 patch 効果の理論的解析を行う．この解析では，画像の中心点に近い pixel ほど，画像中心部分の領域に高い効果を与えることが可能であることを示す．

3.1 一つの pixel からの拡散効果の算出

本小節では，一つの pixel からの拡散効果の算出を行う．はじめに，3.1.1 節において，厳密な拡散効果の算出を行う．次に，算出された拡散効果を近似することで，敵対的 patch の効果が，

- 敵対的 patch の影響は，元々の 1 pixel からの距離に応じて減衰する (3.1.2 節)
- 敵対的 patch の影響は，元々の 1 pixel を中心として，ほぼ同心円状に拡散する (3.1.3 節)

ことを示す．なお，本稿では，画像の外側への拡散による敵対的 patch 効果の減衰は微小なものと仮定し，その減衰を考慮しない議論を行う．これは，本稿で行う理論的解析が画像の中心部分に着目しているからである．

3.1.1 厳密な拡散効果

1 pixel からの敵対的 patch 効果は，以下の通り拡散する．

$$\begin{bmatrix}1\end{bmatrix} \to \frac{1}{9}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix} \to \frac{1}{81}\begin{bmatrix}1&2&3&2&1\\2&4&6&4&2\\3&6&9&6&3\\2&4&6&4&2\\1&2&3&2&1\end{bmatrix} \to \cdots$$

本稿では，n 回の拡散後，patch の拡散範囲を表す行列を

$D(n)$ とする．行列 $D(n)$ について，以下の定理 1 が成立する．

定理 1. 行列 $D(n)$ は $(2n+1)\times(2n+1)$ 行列であり，各要素は，$-n \le x \le n, -n \le y \le n$ の下で，

$$D(n)_{n+|x|,n+|y|} = \frac{f(|x|)f(|y|)}{9^n} \tag{1}$$

$$f(a) = \sum_{i=a}^{\lfloor \frac{n+a}{2} \rfloor} \frac{n!}{(n-2i+a)!i!(i-a)!} \tag{2}$$

で表される．

定理 1 において，(n,n) 要素が元々の pixel であり，拡散の核となる．定理 1 は，以下の二つの補題 2,3 によって証明される．

補題 2. 行列 $D(n)$ の (x,y) 成分について，

$$9^n D(n)_{x,y} = \left(9^n D(n)_{x,0}\right)\left(9^n D(n)_{0,y}\right) \tag{3}$$

が成立する．

補題 3. 行列 $D(n)$ について，$-n \le x \le n$ の下で，$D(n)_{n+x,0} = D(n)_{0,n+x} = 9^{-n} f(|x|)$ が成立する．

以下，補題 2 及び 3 を示す．

補題 2 の証明

Patch 効果の拡散回数 n に関する数学的帰納法により証明する．

(1) $n=0$ のとき，$D(0)_{0,0}=1$ より，成立する．

(2) $n=k$ のとき，式 (3) が成立するとする．ここで，$n=k+1$ とすると，

$$\begin{aligned}
&9^{k+1}D(k+1)_{x,y} \\
&= \sum_{i=-1}^{1}\sum_{j=-1}^{1} 9^k D(k)_{x+i,y+j} \\
&= \sum_{i=-1}^{1}\sum_{j=-1}^{1} \left(9^k D(k)_{x+i,0}\right)\left(9^k D(k)_{0,y+j}\right) \\
&= 9^{2k+2}\left(\frac{1}{9}\sum_{i=-1}^{1} D(k)_{x+i,0}\right)\left(\frac{1}{9}\sum_{j=-1}^{1} D(k)_{0,y+j}\right) \\
&= \left(9^{k+1}D(k+1)_{x,0}\right)\left(9^{k+1}D(k+1)_{0,y}\right)
\end{aligned}$$

より，$n=k+1$ のときも成立する．

以上より，補題 2 が示せた． ■

補題 3 の証明

$0 \le x \le n$ の下で，$D(n)_{n+x,0} = 9^{-n} f(x)$ であることを示す．なお，以上が示せれば，他の状況については，対称性より明らかである．

はじめに，敵対的 patch の拡散方法について考察を行う．Oonishi らのモデルにおいて，敵対的 patch の影響値を，周囲 9 マスの影響値の平均値へ更新している．上記を言いかえると，各 pixel の影響値が，1/9 ずつ周囲 9 マスに拡散していることとなる．したがって，(n,n) を起点とした場合，任意の pixel$(n+x,n+y)$ への影響値は移動方法の集合 P を $P=\{(i,j)\,|\,-1\le i\le 1,-1\le j\le 1\}$ としたとき，

$$D(n)_{n+x,n+y} = \frac{1}{9^n}\#\left(\sum_{k=1}^{n} P_k = (x,y)\,,\text{where } P_k \in P\right)$$

となる．よって，$D(n)_{n+x,0}$ を求めるためには，(n,n) から $(n+x,0)$ への移動方法の個数を数えればよい．

ここで，$(n+x,0)$ は，敵対的 patch の拡散範囲のうち，外枠の pixel(左端の列) を指し示している．したがって，(n,n) から $(n+x,0)$ に至る場合，P_k は常に $(i,-1)$ なる構造を持つ．ゆえに，P_k の第一成分について，その総和が x となるような移動方法の個数を数えればよい．ここで，P_k の第一成分の総和が x となるような n 回の移動方法は，$x \le i \le \left\lfloor \frac{n+x}{2} \right\rfloor$ の下で，

- $(0,-1)$: $(n-2i+x)$ 回
- $(1,-1)$: i 回
- $(-1,-1)$: $i-x$ 回

となり，それぞれの i における移動方法は

$$\frac{n!}{(n-2i+x)!i!(i-x)!}$$

通りとなる．これらの値の i に関する総和は $f(x)$ に等しいため，$D(n)_{n+x,0} = 9^{-n} f(x)$ となる．以上より，補題 3 が示せた． ■

3.1.2 敵対的 patch 効果の減衰に関する評価

本小節では，前小節で得られた敵対的 patch の拡散効果を近似し，敵対的 patch の影響が，元々の 1 pixel からの距離に応じて減衰することを示す．以下の議論では，前小節と同様，$0 \le x \le n$ における議論を行う．

はじめに，式 (2) の近似を行う．スターリングの公式より，$\log n! \approx n\log n - n$ なので，

$$\begin{aligned}
&\log \frac{n!}{(n-2i+x)!i!(i-x)!} \\
&\approx (n\log n - n) - (i\log i - i) \\
&\quad - ((i-x)\log(i-x) - (i-x)) \\
&\quad - ((n-2i+x)\log(n-2i+x) - (n-2i+x)) \\
&= n\log n - i\log i - (i-x)\log(i-x) \\
&\quad - (n-2i+x)\log(n-2i+x)
\end{aligned}$$

となる．ここで，この近似値を $g(i)$ と定義し，$g(i)$ を i で微分すると，

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}i} &= -\log i - \log(i-x) + 2\log(n-2i+x) \\
&= \log\left(\frac{(n-2i+x)^2}{i(i-x)}\right)
\end{aligned}$$

となり，$\frac{\mathrm{d}g}{\mathrm{d}i}$ は，$x \le i \le \left\lfloor \frac{n+x}{2} \right\rfloor$ の下で，i に関して単調

減少である．ここで，$\lim_{i\to x}\frac{\mathrm{d}g}{\mathrm{d}i}\to+\infty$，$\lim_{i\to\lfloor\frac{n+x}{2}\rfloor}\frac{\mathrm{d}g}{\mathrm{d}i}\to-\infty$ であり，

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}i}=0&\Leftrightarrow(n-2i+x)^2=i(i-x)\\&\Leftrightarrow i=\frac{4n+3x-\sqrt{4n^2-3x^2}}{6}\\&\Leftrightarrow i=\frac{4+3(x/n)-\sqrt{4-3(x/n)^2}}{6}n\end{aligned}$$

において，$g(i)$ は最大値を取る．ここで，以上の議論は，指数関数 a^n の底 a に関する最大値を求めている．本稿では，この最大値以外の項を，微小な項として無視する．このとき，$\alpha=x/n$ $(0\le\alpha\le1)$ とし，関数 $A_1(\alpha),A_2(\alpha),$ 及び $A_3(\alpha)$ を

- $A_1(\alpha)=\dfrac{\sqrt{4-3\alpha^2}-1}{3}$
- $A_2(\alpha)=\dfrac{4+3\alpha-\sqrt{4-3\alpha^2}}{6}$
- $A_3(\alpha)=\dfrac{4-3\alpha-\sqrt{4-3\alpha^2}}{6}$

と定義すると，$A_1(\alpha)+A_2(\alpha)+A_3(\alpha)=1$ より，

$$\begin{aligned}f(x)&\approx\frac{n^n}{(A_1(\alpha)n)^{A_1(\alpha)n}(A_2(\alpha)n)^{A_2(\alpha)n}(A_3(\alpha)n)^{A_3(\alpha)n}}\\&=\left(A_1(\alpha)^{A_1(\alpha)}A_2(\alpha)^{A_2(\alpha)}A_3(\alpha)^{A_3(\alpha)}\right)^{-n}\end{aligned}$$

となる．以下，

$$\overline{f}(\alpha)=A_1(\alpha)^{A_1(\alpha)}A_2(\alpha)^{A_2(\alpha)}A_3(\alpha)^{A_3(\alpha)}$$

と定義し，この関数 $\overline{f}(\alpha)$ について以下の定理 4 を示す．定理 4 より，関数 $f(x)$ は x の増大に伴って減衰することが示せる．

定理 4. 関数 $\overline{f}(\alpha)$ は，$0\le\alpha\le1$ における単調増加関数である．

定理 4 の証明

$$\log\overline{f}(\alpha)=\sum_{i=1}^{3}A_i(\alpha)\log A_i(\alpha)$$

であり，関数 $\log\overline{f}(\alpha)$ を α で微分すると，$A_1^2=A_2A_3$，$A_1+A_2+A_3=1$，および $A_2\ge A_3$ より，

$$\begin{aligned}\frac{\mathrm{d}\log\overline{f}(\alpha)}{\mathrm{d}\alpha}&=\sum_{i=1}^{3}\frac{\mathrm{d}A_i(\alpha)}{\mathrm{d}\alpha}+\sum_{i=1}^{3}\log A_i(\alpha)\frac{\mathrm{d}A_i(\alpha)}{\mathrm{d}\alpha}\\&=\frac{1}{2}\log\frac{A_2(\alpha)}{A_3(\alpha)}\ge0\end{aligned}\tag{4}$$

である．したがって，関数 $\overline{f}(\alpha)$ は，$0\le\alpha\le1$ における単調増加関数である．以上より，定理 4 が示せた． ■

3.1.3 敵対的 patch の拡散効果の等高線

本節では，$(0,0)$ の近傍にある (x,y) について，敵対的 patch の拡散効果 $D(n)_{n+x,n+y}$ の等高線の解析を行う．この解析により，$(0,0)$ の近傍では，敵対的 patch の拡散効果 $D(n)_{n+x,n+y}$ の等高線がほぼ同心円となることが示せる．$x\ge0,y\ge0$ の領域において，$D(n)_{n+x,n+y}=9^{-n}f(x)f(y)$ である．この関数 $D(n)_{n+x,n+y}$ について，x で偏微分を行うと，

$$\begin{aligned}\frac{\partial D(n)_{n+x,n+y}}{\partial x}&=9^{-n}f(y)\frac{\mathrm{d}f(x)}{\mathrm{d}x}\\&=9^{-n}f(y)\frac{\mathrm{d}\exp\left(-n\log\overline{f}(\alpha)\right)}{\mathrm{d}\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}x}\\&=-\frac{n9^{-n}f(y)}{n}\frac{\mathrm{d}\log\overline{f}(\alpha)}{\mathrm{d}\alpha}f(x)\\&=-\frac{1}{2}\log\frac{A_2(\alpha)}{A_3(\alpha)}D(n)_{n+x,n+y}\ (\because\text{式 (4)})\end{aligned}$$

である．y に関する偏微分も同様なので，$\beta=y/n$ $(0\le\beta\le1)$ とすると，関数 $D(n)_{n+x,n+y}$ の勾配は，

$$-\frac{1}{2}D(n)_{n+x,n+y}\left[\log\frac{A_2(\alpha)}{A_3(\alpha)},\log\frac{A_2(\beta)}{A_3(\beta)}\right]\tag{5}$$

となる．ここで，$0\le\alpha\le1$ において，

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\log\frac{A_2(\alpha)}{A_3(\alpha)}\right)&=\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\log A_2(\alpha)-\log A_3(\alpha)\right)\\&=\frac{1}{A_2(\alpha)}\frac{\mathrm{d}A_2(\alpha)}{\mathrm{d}\alpha}-\frac{1}{A_3(\alpha)}\frac{\mathrm{d}A_3(\alpha)}{\mathrm{d}\alpha}\end{aligned}$$

であり，

$$\lim_{\alpha\to0}A_1(\alpha)=\lim_{\alpha\to0}A_2(\alpha)=\lim_{\alpha\to0}A_3(\alpha)=\frac{1}{3},$$

$$\lim_{\alpha\to0}\frac{\mathrm{d}A_2(\alpha)}{\mathrm{d}\alpha}=\lim_{\alpha\to0}\left(\frac{1}{2}+\frac{\alpha}{2\sqrt{4-3\alpha^2}}\right)=\frac{1}{2},$$

$$\lim_{\alpha\to0}\frac{\mathrm{d}A_3(\alpha)}{\mathrm{d}\alpha}=\lim_{\alpha\to0}\left(-\frac{1}{2}+\frac{\alpha}{2\sqrt{4-3\alpha^2}}\right)=-\frac{1}{2}$$

より，$(0,0)$ の近傍にある (x,y) について，

$$D(n)_{n+x,n+y}\sim1,\log\frac{A_2(\alpha)}{A_3(\alpha)}\sim3\alpha$$

となる．したがって，$(0,0)$ の近傍にある (x,y) における勾配は，$(-3x/(2n),-3y/(2n))$ となるので，敵対的 patch の拡散効果 $D(n)$ の等高線はほぼ同心円となる．

3.2 YOLOv2 [7] 上での敵対的 patch 効果の理論的解析

3.1.3 節で，一つの pixel からの拡散効果は，pixel の近傍において同心円で近似できることを確認した．本小節では，YOLOv2 において，指定された領域に対する効果が高くなるような敵対的 patch の配置法について理論的解析を行う．本小節では，ある一つの pixel からの拡散効果について以下の仮定を置き，理論的解析を行う．

仮定 5. 各 pixel の効果は，同心円状に拡散する．特に，起点となる pixel から (x,y) にある点における勾配は，$(-3x/(2n),-3y/(2n))$ とする．

ここで，以上の仮定 5 は，一見，YOLOv2 のプーリング層を無視しているように見える．しかし，実際には，拡散

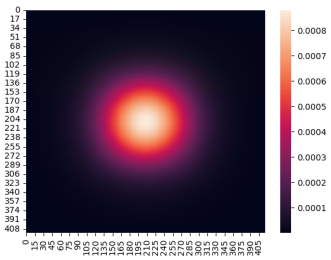

図 1 中心の 96×96 の領域 R に対する各 pixel の影響値
Fig. 1 Effect values on the central 96×96 area R

図 2 中心の 96×96 の領域 R に対して強力な遠隔敵対的 patch
Fig. 2 Strong remote adversarial patches for the central 96×96 area R

効果の計算は，各 pixel に対して線形な演算である．よって，YOLOv2 のプーリング層では，画像サイズを保持し，pixel の総和を取らなければ，

- pixel 効果の拡散が x マス単位から $2x$ 単位に変化
- 隣接する pixel は，独立に拡散を行う

のように演算ルールが変化した，とみなすことができる．したがって，YOLOv2 のプーリング層を無視して，以上の仮定 5 によって近似する．また，以上の仮定 5 は，$(x, y) = (0, 0)$ の近傍の近似を全体に拡張しているが，この近似の精密な解析は今後の課題である．以上より，本小節では，仮定 5 の下で，指定された領域に対する効果が高くなるような敵対的 patch の配置法について理論的解析を与える．

以下，本稿では，画像中心の 96×96 サイズの領域 R (実際には，最終層における 3×3 サイズの領域) に対する敵対的 patch の効果を最大化するような配置法について考察する．なお，この領域 R は，$R = \{(a, b) \,|\, 156 \leq a \leq 251, 156 \leq b \leq 251\}$ である．ここで，$(208, 208)$ を起点として (x, y) にある点 $(208 + x, 208 + y)$ が領域 R に与える影響値を $F(x, y)$ とする．このとき，(x, y) について，$F(x, y)$ の勾配は，$\sum_{i=-48}^{48} \left(-\frac{3}{2n}(x - i), -\frac{3}{2n}(y - i) \right) = \left(-\frac{72}{n}x, -\frac{72}{n}y \right)$ となる．よって，$F(x, y)$ を一定値とするような等高線は，$(208, 208)$ を中心とする同心円となる．したがって，敵対的 patch の効果を最大化するためには，$(208, 208)$ からの距離が近い方から，pixel を選択すればよい．

ここで，本研究では，以上を検証するため，拡散モデルを利用した検証を行った．本実験では，Oonishi らの拡散モデルを利用し，各 pixel が領域 R に与える影響値の総和を計算し，その影響値を図示する．図示した結果は図 1 の通りである．上記の解析の通り，$(208, 208)$ からの距離が近い pixel の効果が高くなっているとともに，その等高線は，$(208, 208)$ を中心とする同心円となっている．

4. 研究成果 2: 強力な遠隔敵対的 patch の生成アルゴリズムの提案

本節では，敵対的 patch の拡散モデルを利用した，強力な遠隔敵対的 patch の生成アルゴリズムの提案を行う．本稿で提案するアルゴリズムは，敵対的 patch の配置方法の計算 (4.1 節)，敵対的 patch の生成 (4.2 節) の 2 段階によって敵対的 patch 生成を行う．

4.1 敵対的 patch の配置方法の計算

まず，第一段階目の，敵対的 patch の配置方法の計算では，どの位置に敵対的 patch を配置すれば良いか計算を行う．具体的には，以下によって，攻撃能力の高い敵対的 patch の配置法を計算する．

(1) 敵対的 patch の効果を与える領域を指定する．
(2) 各 pixel について，指定領域に与える影響値の合計値を計算する．
(3) 敵対的 patch 配置予定の領域について，影響値の合計値が大きいほうから，指定された数になるまで pixel を選択する．

ここで，本稿では，前節同様，敵対的 patch の効果を与える領域は画像中心の 96×96 なる領域 R とする．また，敵対的 patch 配置予定の領域は，既存研究 [6] と同様に，$S = \{(x, y) \,|\, (0 \leq x \leq 104 \vee 311 \leq x \leq 415) \wedge 0 \leq y \leq 415\}$ で定義される領域 S とする．

以上の計算方法により導出した，領域 R に対して強力な遠隔敵対的 patch は，図 2 の通りである．図 2 の敵対的 patch は，$(208, 208)$ を中心とする円と領域 S の共通部分であるといえる．実際，左側の敵対的 patch 部分の弧は，$(103, 109)$, $(70, 208)$, 及び $(103, 305)$ の三点を通っている．それぞれ，$(208, 208)$ からの距離は，144.3，138，及び 147.1 となるので，少し円よりは広がった形状となっているものの，大体同心円とみなしてよい．したがって，前節の理論的解析は正当である．

4.2 敵対的 patch による物体検出率の変化

次に，以上で計算した敵対的 patch の領域において，敵

対的 patch の生成を行う．本実験では，Oonishi らの既存研究 [6] と同じく，Saha らの実装 [8] に基づいて敵対的 patch を生成する．具体的には，図 2 の配置方法により，全ての RGB 値が 0 である状態から学習を開始し，

$$(\text{クラス値}) + 0.01 \times (\text{NPS})$$

を最小化することで敵対的 patch を生成する．なお，NPS 項の倍率は，Thys らの実装 [10] に基づいて決定した．また，本稿の RGB 値は，通常 256 段階ある RGB 値を，0 から 1 の範囲にマッピングした値である．その上で，生成した敵対的 patch を配置した際の平均検出率を比較する．なお，平均検出率は，

(1) 各画像について bounding box をすべて走査し，最大の検出率を導出
(2) (1) で得られた検出率の平均を計算

によって計算した．

実験結果は，表 2 の通りであるなお，表 2 の **average** は，表 2 にある各クラスの検出率の加算平均を取った値である．本研究で生成した敵対的 patch は，既存研究と比較した際，検出率が改良していないクラスはあるものの，平均検出率は 12.81%から 9.27%に低下している．特に，本研究で作成した遠隔敵対的 patch は，既存研究で 10%の検出率を超えるクラスについては，検出率を下げることに成功している．したがって，本研究で作成した遠隔敵対的 patch は，既存研究の遠隔敵対的 patch よりも強力である．

5. 研究成果 3: 動的な遠隔敵対的 patch の生成実験

前節までは，Oonishi らが提案した敵対的 patch 効果の拡散モデルに基づいた理論的解析を行い，最適な敵対的 patch の配置法について理論的な解析を行ってきた．特に，前節までの議論は，モデルから算出された計算値により，最適な敵対的 patch の配置法を事前に定めることを目的としてきた．本節では，前節までとは異なり，Huang ら [4] や Zhu ら [12] の手法と同様，事前に敵対的 patch の位置を固定せず，最適な遠隔敵対的 patch の形状を学習する．

本稿では，最適化関数について，前節でのクラス値，NPS 項に加え，各 pixel の RGB 値の 2-norm の総和を加えた値 (pixel 値) を最小化する．特に，本稿では，前節の実験とは異なり，全ての RGB 値が 1 である状態から学習を開始し，

$$(\text{クラス値}) + 0.01 \times (\text{NPS 項}) + 10^{-5} \times (\text{pixel 値})$$

を最小化する．全ての RGB 値を 0 として開始しなかったのは，後述する閾値により，敵対的 patch の RGB 値が 0 のまま学習が進行するのを防止するためである．また，pixel 値の倍率は，

- 大体 10%程度になるクラス値と同程度にする
- 0 でない pixel 数が $O\left(10^4\right)$ 個となるように設定する

の二点を満たすよう，$0.1/10000 = 10^{-5}$ として設定した．上記に加え，本実験では，敵対的 patch の内，影響度合いの小さい部分を削減するため，ある閾値を定め，閾値以下となった pixel の RGB 値を 0 とした．具体的には，閾値 t を定めた上で，t 以下となった RGB 値を 0 として，敵対的 patch の生成を行った．

本稿では, 以上の学習方法を VOC データセットの person クラスに適用し，$t = 0.01, 0.02$ の下で実験を行った．これらの値は，256 段階の離散的な RGB 値のうち，それぞれ 2 段階，5 段階の微小な変化を示す．$t = 0.01, 0.02$ について，0 でない RGB 値を持つ pixel 値は，それぞれ，58,401 個，45,019 個となった．$t = 0.01$ 及び 0.02 の場合に生成された遠隔敵対的 patch は，それぞれ図 3，4 の通りである．図 3，4 より，遠隔敵対的 patch は，主に，中心付近に集まっており，その形状は画像の中央を中心とする円の一部となる．以上より，動的に遠隔敵対的 patch の形状を学習しても，その形状は前節までの理論的解析の結果と一致する．したがって，前節までの理論的解析は正当である．

表 2 同面積の敵対的 patch における検出率の変化
Table 2 Detection rates on adversarial patches with the same area

	既存研究 [6]	本研究
枚数	2 枚	2 枚
面積	9,800	9,800
aeroplane	4.04%	5.45%
bicycle	5.94%	7.00%
bird	4.78%	6.20%
boat	3.09%	3.10%
bottle	29.17%	21.35%
bus	1.55%	0.45%
car	18.93%	6.85%
cat	2.55%	1.59%
chair	21.18%	15.07%
cow	7.15%	9.03%
diningtable	45.53%	38.70%
dog	5.81%	4.78%
horse	0.59%	0.05%
motorbike	17.33%	9.27%
person	26.20%	12.71%
pottedplant	23.88%	21.09%
sheep	8.21%	8.92%
sofa	3.51%	1.45%
train	2.60%	1.29%
tvmonitor	24.20%	11.11%
average	12.81%	**9.27%**

6. 結論

本稿では，敵対的 patch の脅威を定量的に見積もるため，Oonishi らの拡散モデル [6] に基づき，

- 敵対的 patch 効果の拡散の近似的な理論的解析

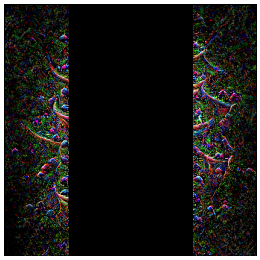

図 3 $t = 0.01$ の下で動的に生成された遠隔敵対的 patch
Fig. 3 Remote adversarial patches generated dynamically when $t = 0.01$

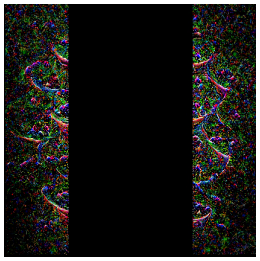

図 4 $t = 0.02$ の下で動的に生成された遠隔敵対的 patch
Fig. 4 Remote adversarial patches generated dynamically when $t = 0.02$

- 最適な遠隔敵対的 patch の配置アルゴリズムの提案
- 動的な遠隔敵対的 patch の生成による理論的解析の実験的検証

を行った．本稿では，特に，画像の中心部分にある物体に対する攻撃を想定した理論的解析を行った．

本稿では，まず，敵対的 patch 効果の拡散について近似的な理論的解析を行い，敵対的 patch の効果がほぼ同心円状に拡散することを示した．その後，拡散モデルに基づいた最適な遠隔敵対的 patch の配置アルゴリズムの提案を行い，遠隔敵対的 patch の形状が理論的解析と同じく円の一部であることを示し，既存研究の遠隔敵対的 patch と比較して，検出率が低下することを示した．さらに，動的な遠隔敵対的 patch 生成においても，理論的解析と同じく，その形状は円の一部となることを示した．

本稿では，解析のため，厳密な評価を行った定理 1 に基づき，様々な近似を行って，敵対的 patch 効果の評価を行った．しかし，これらの近似が完全に妥当かどうか，については今後，さらなる厳密な解析が必要である．特に，該当 pixel の近傍における近似を全体に拡張しているため，実際のはどの程度の差異が生じるかの精密な解析は今後の課題である．

さらに，本稿では，画像の外側への拡散による patch 効果の減衰を無視した議論を行った．本稿では，画像の中心付近の領域に対する議論を行ったため，上記の近似は妥当であると考えられる．だが，上記の近似が，実際の敵対的 patch 効果にどの程度の影響を及ぼすかの精密な解析は今後の課題である．

また，本稿では，画像の中心にある 96×96 サイズの画素を踏まえた敵対的 patch の生成を行っている．今後，画像のどの部分を踏まえて敵対的 patch の生成を行えば良いかを考察することも今後の課題である．

最後に，本稿の敵対的 patch をより現実的な脅威として見積もるため，YOLOv2 以外の CNN に適用すること及び物理的な敵対的 patch を実現することも今後の課題である．

謝辞

本研究の成果は国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）の結果得られたものである．

参考文献

[1] Brown, T. B., Mané D., Roy, A., Abadi, M., and Gilmer, J.: Adversarial Patch, eprint arXiv 1712.09665 (2017).

[2] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective, International Journal of Computer Vision, 111(1), 98–136 (2015).

[3] Hoory, S., Shapira, T., Shabtai, S., and Elovici, Y.: Dynamic Adversarial Patch for Evading Object Detection Models, eprint arXiv 2010.13070 (2020).

[4] Huang, H. Wang, Y., Chen, Z., Tang, Z., Zhang, W., and Ma, K.-K.: RPAttack: Refined Patch Attack on General Object Detectors, eprint arXiv 2103.12469 (2021).

[5] Liu, X., Yang, H., Liu, Z., Song, L., Li, H., and Chen, Y.: DPatch: An Adversarial Patch Attack on Object Detectors, eprint arXiv 1806.02299 (2018).

[6] 大西健斗，中井綱人，鈴木大輔: 物体検出 CNN に対する複数配置に着目した遠隔 Adversarial Patch 攻撃，SCIS 2022, 3A2-4 (2022).

[7] Redmon, J. and Farhadi, A.: YOLO9000: Better, Faster, Stronger, CVPR 2017, pp. 6517–6525 (2017).

[8] Saha, A., Subramanya, A., Patil, K., and Pirsiavash, H.: Role of Spatial Context in Adversarial Robustness for Object Detection, CVPRW 2020, pp.3403–3412 (2020). Avaliable source code at `https://github.com/UMBCvision/Contextual-Adversarial-Patches`

[9] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K.: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, CCS2016, pp. 1528–1540 (2016).

[10] Thys, S., Ranst, W.V., and Goedemé, T.: Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection, CVPRW 2019, pp. 49–55 (2019). Avaliable source code at `https://gitlab.com/EAVISE/adversarial-yolo`

[11] Yu, C., Chen, J., Xue, Y., Liu, Y., Wan, W., Bao, J., and Ma, H.: Defending against Universal Adversarial Patches by Clipping Feature Norms, ICCV 2021, pp. 16414–16422 (2021).

[12] Zhu, Z., Su, H., Liu, C., Xiang, W., and Zheng, S.: You Cannot Easily Catch Me: A Low-Detectable Adversarial Patch for Object Detectors, eprint arXiv 2109.15177 (2021).