Journal of the KIMST, Vol. 26, No. 1, pp. 44-53, 2023
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2023.26.1.044

Research Paper 정보·통신 부문

객체탐지 모델에 대한 위장형 적대적 패치 공격

김정훈[1] · 양훈민[1] · 오세윤*,[1]

[1] 국방과학연구소 국방첨단과학기술연구원

Camouflaged Adversarial Patch Attack on Object Detector

Jeonghun Kim[1] · Hunmin Yang[1] · Se-Yoon Oh*,[1]

[1] *Advanced Defense Science & Technology Research Institute, Agency for Defense Development, Korea*

(Received 28 October 2022 / Revised 26 January 2023 / Accepted 31 January 2023)

Abstract

Adversarial attacks have received great attentions for their capacity to distract state-of-the-art neural networks by modifying objects in physical domain. Patch-based attack especially have got much attention for its optimization effectiveness and feasible adaptation to any objects to attack neural network-based object detectors. However, despite their strong attack performance, generated patches are strongly perceptible for humans, violating the fundamental assumption of adversarial examples. In this paper, we propose a camouflaged adversarial patch optimization method using military camouflage assessment metrics for naturalistic patch attacks. We also investigate camouflaged attack loss functions, applications of various camouflaged patches on army tank images, and validate the proposed approach with extensive experiments attacking Yolov5 detection model. Our methods produce more natural and realistic looking camouflaged patches while achieving competitive performance.

Key Words : Adversarial Attack(적대적 공격), Camouflage(위장), Deep Learning(딥러닝), Machine Learning(머신러닝)

기 호 설 명

L : Loss function
c : Comparable image pixel value
b : Image pixel value
MP : Military pattern
BG : Background

* Corresponding author, E-mail: syoh@add.re.kr

1. 서 론

최근 인공신경망을 활용한 이미지 분류, 객체탐지 기술이 보편화되고 자율주행 자동차, 로봇 팔 제어와 같은 다양한 분야에 해당 기술이 적극적으로 적용됨에 따라 이러한 인공신경망 기반 이미지처리 기능들의 성능저해를 위한 적대적 공격의 위험 속에 노출되어 있다. 일반적으로 적대적 공격이란 인공신경망을 통한 이미지 분류, 객체 탐지 등의 성능을 인위적으로 감소시키기 위한 모든 방법을 지칭한다[1]. 이 중 상대

방이 사용하는 인공신경망 시스템에 대한 접근이 제한되는 국방분야에서는 입력 데이터를 조작하는 방식의 회피(Evasion)공격이 효과적으로 사용될 수 있다.

회피 방식의 적대적 공격은 실물에 가해지는 조작 및 변형의 종류에 따라 적대적 조명, 적대적 형상, 적대적 패치 공격 등으로 구분되며 가장 널리 사용되는 방식인 적대적 패치 방법은 공격 대상 객체에 특정한 교란 이미지를 부착하거나 표면 전체에 적용하는 공격 기법을 말한다. 도로 표지판이나 위성사진 속 비행기와 같이 관측되는 시야가 한정적인 경우에는 물체보다 작은 패치를 부착[2][3]하기도 하며, 도로 위의 자동차와 같이 모든 방향에서 관측되는 상황에서는 객체의 표면 전체에 최적화된 패치를 적용하는 방식[4,5]으로 인공신경망 기반탐지능력을 무력화시킨다.

적대적 조명 공격은 공격 대상 물체에 레이저[6], 빔 프로젝터[7], LED[8] 등의 광원을 통한 공격기법을 통칭한다. 이러한 적대적 조명 공격기법은 조명을 조절하거나 적외선과 같이 관측이 제한적인 조명을 사용하여 은밀성 높은 적대적 공격을 수행할 수 있다는 장점을 가지지만, 공격이 특정 시야각에 한정되고 외부 광원에 의해 무력화될 수 있다는 단점이 있다. 적대적 형상 공격기법은 공격 대상 물체에 추가적인 3차원 형상을 부착하거나[9] 물체의 형상 자체를 조작하는 방식[10,11]의 공격기법을 말한다. 라이더 감지기를 공격하기 위한 목적으로 주로 사용되지만, 공격을 위한 최적화가 어렵고 실물 적용이 복잡하다는 단점이 있다.

디지털 공격 이외의 모든 종류의 적대적 공격은 공격성능의 향상을 위해 공격 대상 물체에 높은 수준의 교란 및 변형을 가하는 경우가 많다. 이러한 큰 교란과 변형은 적대적 공격을 통해 인공신경망을 효과적으로 기만할 수 있지만 물체에 대한 교란효과가 사람에게 보다 더 잘 인지되도록 하는 단점을 갖게 되어 은밀성이 요구되는 국방분야 적용에는 기존의 적대적 공격기법들의 적용에 큰 제약이 뒤따른다. 그간 국방분야에서는 사람의 인지를 기만하기 위한 위장기술이 주로 연구되어 왔으며 다양한 위장패턴을 개발하고 그 위장성능을 평가하는 기술이 발전해 왔다. 본 연구에서는 이러한 위장성능 평가지표들을 적대적 패치 최적화 과정의 손실함수에 반영시킴으로써 공격성능과 위장성능 모두를 만족시키는 적대적 패치 최적화를 수행하고자 하였다. 동일한 전차 사진에 대해 위장에 대한 고려가 없이 생성된 적대적 패치와 위장성능을 고려하여 각기 다른 방법으로 최적화된 적대적 패치 적용

(a) Normal patch

(b) Camouflaged patch

Fig. 1. Visualization of normal perceptible patch and our camouflaged naturalistic patches

결과를 Fig. 1에 나타내었다. Fig. 1 (a)의 적대적 패치가 부착된 전차는 육안으로 쉽게 식별될 수 있고, 이는 적대적 패치를 통한 공격 방법이 국방 분야에 적용되기 위해서는 Fig. 1 (b)의 적대적 패치들과 같이 위장성능이 고려된 적대적 패치가 필요함을 의미한다.

본 연구에서는 적대적 공격기술과 국방분야의 위장성능 평가기술에 대한 기술분석을 통해 인공신경망과 사람의 인지를 동시에 회피할 수 있는 적대적 공격

방법론에 대한 고찰을 수행하고 적대적 공격을 위한 최적화와 실물적용이 용이한 장점을 갖고 있는 적대적 패치 부착 공격방식에 대한 유효성을 검증하고자 하였다.

2. 적대적 위장형 패치 최적화 방법

본 장에서는 디지털 이미지 상에서 적대적 패치가 실물적용되는 과정을 모사하여 구현하고, 이를 바탕으로 효율적인 적대적 패치 최적화를 수행하기 위한 학습데이터, 객체탐지 모델, 그리고 손실함수 등을 설정하였다. 또한 국방분야에서 사용되는 위장패턴 및 위장성능 평가 지표들을 적용함으로서 기존의 적대적 패치 최적화 과정을 개선한 위장형 패치 최적화 방법론들을 제시하고, 그 장단점 및 특성들에 대해 논하였다. 모든 연산은 2080 Ti GPU 4개가 장착된 단일 워크스테이션에서 수행하였다.

2.1 위장성능 평가 지표

국방분야에서 위장이란 적군의 탐지로부터 아군의 인적, 물적 자원을 보호하는 모든 행위를 뜻한다. 보다 넓은 의미로는 군사적 목적으로 사용되는 차량, 선박, 비행체, 화기, 군복 등에 색상 및 물질적인 조작을 가하여 적의 탐지로부터 은닉하거나 본래의 물체가 아니라 다른 물체로 인지되도록 하는 모든 행위를 포괄한다. 현재 국군에서 사용되는 위장패턴의 경우 기본적으로 우리나라의 지형과 색상을 고려하여 흑색, 녹색, 갈색, 모래색을 위장패턴의 색상으로 선정하고, 그 형태는 미군의 우드랜드 패턴 및 디지털 패턴을 중심으로 독자적인 패턴을 개발했다. 국방분야에서 활용되는 위장패턴의 위장성능을 평가하기 위한 연구는 위장패턴을 사용한 20세기 초부터 지속적으로 이루어져왔다. 그러나 아직까지 위장성능을 완벽하게 평가할 수 있는 표준적인 군용 위장성능 평가 방식은 없으며, 다양하게 제시된 위장성능 평가 지표들이 복합적으로 사용되고 있다[12]. 다양한 군용 위장성능 평가 방식들은 크게 관측자의 주관적 판단과 대규모 설문조사에 기반을 둔 통계적 평가방식[13], 위장이 적용된 실물을 라이더나 레이더 와 같은 다양한 감지기들로 관측하여 어지는 물리적인 측정값을 기반으로 한 수식적 평가 방식[14], 그리고 위장이 적용된 물체의 사진 및 동영상을 기반으로 이미지 픽셀의 객관적 수치에 기반을 둔 수식적 평가방식 세 가지로 구분할 수 있다.

군용 위장패턴 성능평가 기법을 적대적 공격을 위한 적대적 패치 최적화에 적용하기 위해서는 위장성능에 대한 지표가 손실함수의 형태로 계산되어 연속적인 최적화 과정에 포함될 수 있어야한다. 앞서 제시된 세 종류의 방법들 중 이에 가장 적합한 평가방식은 이미지 픽셀의 객관적 수치에 기반을 둔 수식적 평가방식이다. 이미지 자체가 연산 및 최적화가 가능한 데이터의 형태이기 때문에 손실함수에 따라 데이터를 변조하는 일반적인 학습 방법론을 그대로 적용할 수 있다. 이러한 적대적 공격 최적화 루프에 적용되기에 가장 적합한 3가지의 위장성능 평가지표들인 MSE(Mean Square Error)[12], PSNR(Peak Signal to Noise Ratio)[12], UIQI(Universal Image Quality Index)[12]을 각각 수식 (1)~(3)에 나타내었다.

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(b_{ij}-c_{ij})^2 \tag{1}$$

$$PSNR = 10\log_{10}\left(\frac{D^2MN}{\sum_{i=1}^{M}\sum_{j=1}^{N}(b_{ij}-c_{ij})^2}\right) \tag{2}$$

$$UIQI = \frac{\sigma_{bc}}{\sigma_b\sigma_c}\frac{2\bar{b}\bar{c}}{\bar{b}^2+\bar{c}^2}\frac{2\sigma_b\sigma_c}{\sigma_b^2+\sigma_c^2} = \frac{4\bar{b}\bar{c}\,\sigma_{bc}}{(\bar{b}^2+\bar{c}^2)(\sigma_b^2+\sigma_c^2)} \tag{3}$$

이 외에도 GabRat(Gabor edge disruption ratio), 주목성 분포도(Saliency map) 등의 위장성능 평가지표들도 손실함수의 형태로 변환될 수 있으나, 알고리즘의 효율성 및 범용적인 지표 적용을 위해 본 논문에서는 위 세 지표들을 바탕으로 위장형 적대적 공격 패치를 최적화하였다. 주어진 이미지에 대해 R, G, B 세 채널들에 대해 각각 MSE, PSNR, UIQI를 계산하고 동등한 비율로 합한 값을 손실함수로 사용했고 UIQI의 경우에는 0과 1 사이의 값을 가지며 1에 가까울수록 위장성능이 높은 것이기에 (1-UIQI)의 값을 최소화했다.

2.2 적대적 패치 최적화

적대적 공격 성능 극대화를 위한 손실함수 L_{adv}는 세부적으로 L_{obj}, L_{nps}, L_{tv}의 세 세부 손실함수들로 구분된다. 각 손실함수들은 적대적 공격을 효과적으로 수행하기 위해 최적화되는 세부목적들을 위한 손실함

수이고, L_{adv}는 이들의 선형결합으로 표현된다. L_{obj}, L_{nps}, L_{tv}의 선형결합 시 가중치는 해당 손실함수의 목적을 강조하기 위해 다른 손실함수들의 값이 무시되지 않는 적절한 범위 내에서 조절될 수 있다.

L_{obj}는 적대적 패치가 객체 중앙에 부착된 이미지들에 대한 Yolov5s 객체탐지모델의 최대 객체지수(Objectness score) 값으로 정의되며 수식 (4)[15]와 같이 표현된다. 이 손실함수의 값이 작을수록 객체탐지 모델이 이미지 내에서 객체의 존재를 잘 찾지 못하는 것이므로, 이를 최소화하도록 적대적 패치를 최적화함으로써 공격 성능을 극대화할 수 있다.

$$L_{obj} = \max_k p_k \tag{4}$$

L_{tv}는 픽셀 단위로 색을 자유롭게 선택 가능한 디지털 데이터 영역과는 달리 적대적 패치가 실물에 적용될 때 발생하는 왜곡 및 한계로 인한 잠재적인 공격 성능 저하를 최소화하는 함수이며 수식 (5)[15]와 같이 표현된다. 이 손실함수의 값은 각 픽셀의 색상 값들과 인접한 픽셀들의 색상의 값들과의 분산의 총합으로 정의된다.

$$L_{tv} = \sum_{i,j} \sqrt{(b_{i,j} - b_{i,j+1})^2 + (b_{i,j} - b_{i+1,j})^2} \tag{5}$$

L_{nps}는 디지털 데이터 영역에서 최적화되는 적대적 패치가 실물에 적용될 때 디지털 데이터로는 구현되는 색상이 실제로는 구현이 불가능하여 발생하는 적대적 패치의 잠재적인 공격 성능 저하를 최소화하며 수식 (6)[15]과 같이 표현된다. 이 손실함수의 값은 적대적 패치의 각 픽셀들의 색상을 사전에 선정된 '실물 적용이 가능한 색상 목록'의 색상들 중 가장 유사한 색상과의 차이의 총합으로 정의된다. 본 연구에서는 L_{nps}를 사용하여 사람 탐지 인공지능 모델을 공격한 유사 연구[15]에서 정의한 '실물 적용이 가능한 색상 목록'을 동일하게 사용하였다.

$$L_{nps} = \sum_{i,j} \min_{c \in C_{print}} |b_{ij} - c| \tag{6}$$

위장성능의 극대화를 위한 손실함수 $L_{camouflage}$은 위장성능을 수치화한 세 종류의 손실함수들로 구성된다. 각각의 세부 손실함수들 L_{MSE}, L_{PSNR}, L_{UIQI}는 위장성능평가 지표를 R, G, B 세 채널들에 대해 동등하게 가중한 값이며, MSE와 PSNR은 값이 작을수록, UIQI는 값이 1에 가까울수록 위장성능이 높음을 의미하기에 각각의 채널들의 MSE, PSNR, (1-UIQI)의 합으로 표현된다. 위장 손실함수는 세부 손실함수들의 선형결합으로 표현되고, 총 손실함수는 적대적 공격 손실함수와 위장 손실함수의 선형 결합으로 표현된다. 각 가중치 α, β, γ들은 L_{obj}, L_{tv}, L_{nps}의 값들이 무시되지 않는 적절한 범위 내에서 조절될 수 있다.

적대적 패치를 최적화하기 위한 총 손실함수 L은 적대적 공격 성능을 위한 손실함수 L_{adv}와 위장성능 극대화를 위한 손실함수 $L_{camouflage}$의 선형결합으로 표현되고 수식 (7)과 같이 표현된다. 기존의 위장을 고려하지 않는 적대적 패치 최적화 과정에서는 L_{obj}, L_{tv}, L_{nps}만을 손실함수로 활용하는 방식이며, 본 논문에서는 위장 성능을 높이기 위하여 $L_{camouflage}$를 추가로 도입하고 L_{adv}과의 선형결합을 손실함수로 활용하는 최적화를 진행하였다.

$$\begin{aligned} L &= L_{adv} + \theta L_{camouflage} \\ &= (\lambda_{obj} L_{obj} + \lambda_{tv} L_{tv} + \lambda_{nps} L_{nps}) \\ &\quad + \theta (\alpha L_{MSE} + \beta L_{PSNR} + \gamma L_{UIQI}) \end{aligned} \tag{7}$$

Fig. 2 (a)의 배경 중심 위장형 적대적 최적화 경로(Background centric camouflaged optimization)와 같이 배경과 적대적 패치 사이의 고준위 특성 간 거리를 최소화하여 는 경우에는 일반적인 전차 이미지와 적대적 패치가 부착된 전차 이미지에 대해, 군용 패턴 중심 위장형 적대적 최적화 경로(Military pattern centric camouflaged optimization)와 같이 군용 위장패턴와 적대적 패치 사이의 시각적 거리를 최소화하는 경우에는 군용 위장패턴과 적대적 패치에 대해 위장성능평가 지표들을 계산하여 손실함수로 반영하였다.

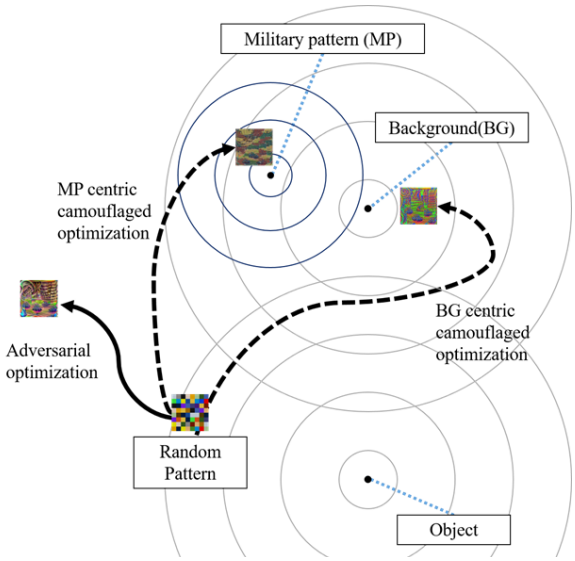
2.3 위장형 적대적 패치 최적화

사람의 인지에 어떠한 물체가 인지되는 것은 그 물체가 주변의 배경과 상이하기 때문이다. 따라서 적대적 공격을 통해 최적화되는 적대적 패치가 사람의 인지에 대한 은밀성을 확보하기 위해서는 적대적 공격 대상 물체가 주로 존재하는 배경과 시각적으로 유사하여야 한다. 국군에서 군용 목적으로 사용되는 위장

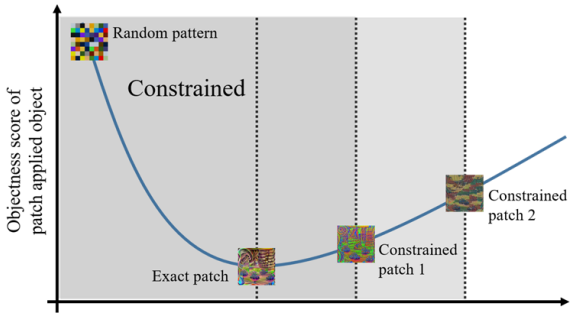
패턴의 경우 전차나 장갑차와 같은 군용 물체가 주로 관측되는 배경인 수풀, 숲 등과 유사한 색상과 무늬를 통해 시각적 유사성을 극대화한 것이라 볼 수 있다. 사람의 인지가 시각적 유사성을 기반으로 물체를 구분하는 반면, 객체탐지 모델은 고준위 특성(Feature) 의 차이로 객체와 배경을 구분한다. 시각적 유사성을 수치화한 위장성능평가 지표들을 손실함수로 사용하여 고준위 특성 차이를 최적화할 수 있다.[16] 이 때 적대적 패치를 통한 공격은 물체에 적용되는 패턴이 물체(Object)와의 고준위 특성거리(High level feature distance)가 멀어질수록, 그리고 적대적 패치가 부착된 객체의 객체점수(Objectness score)가 감소할수록 인공지능이 잘못된 판단을 할 확률이 증가하게 된다. 위장을 고려하지 않고 물체와의 특성거리만을 최대화하는 적대적 최적화(Adversarial optimization) 경로와, 적대적 공격 성능과 위장 효과를 종합적으로 고려하는 두 종류의 적대적 패치 최적화(Camouflaged optimization) 경로들을 Fig. 2 (a)에 나타내었고, 그 결과 생성된 적대적 패치들이 부착된 물체의 객체지수 곡선을 Fig. 2 (b)에 표현하였다. 객체지수에 영향을 주지 않는 무작위 패턴과 비교하여 위장을 고려하지 않는 적대적 패치가 객체지수의 감소폭이 가장 크며, 배경 및 군용 무늬와 유사하도록 최적화 과정을 제한하여 생성한 위장형 적대적 패치는 위장 성능을 높이는 대신 객체지수의 감소폭은 다소 낮다.

적대적 공격에 따른 패치 최적화 과정은 무작위 패치로부터 시작하여 인공신경망에 대한 공격 성능이 높은 패치로 변화하는 과정이다. 이 때 임의의 패치의 적대적 공격 성능은 고준위 특성 공간에서 그 패치와 물체 간 거리가 멀어질수록 증가하기에 Fig. 2 (a)의 위장을 고려하지 않은 적대적 패치 최적화 과정은 무작위 패턴으로부터 물체와 멀어지는 방향으로 이루어진다. 그 결과 적대적 공격 성능은 증대될지언정 최적화 과정을 거치며 더욱 비현실적인 패치로 변화하게 되고, 이로 인해 위장을 고려하지 않은 적대적 공격 패치들은 사람의 인지에는 쉽게 노출된다. 자연적인 배경과 크게 상이한 형상일수록 객체탐지 모델과 같은 인공신경망의 학습 데이터에 포함되지 않는 종류의 형상일 확률이 높기에 공격 성능을 높이는 최적화 과정에서 이러한 경향성은 더욱 두드러진다.

실제 환경에서 이미지 상으로 나타날 수 있는 배경의 다양성이 매우 높기 때문에 모든 현실적인 패치에 대한 종합적인 혹은 평균적인 '배경과의 고준위 특성



(a) High level feature distance optimization



(b) Objectness score minimization

Fig. 2. Patch optimization from random pattern

거리'를 도출하여 최적화 과정에 반영하는 것은 어렵다. 따라서 적대적 공격의 목표를 국방분야로 한정하여 물체의 배경을 대한민국의 자연 환경이라는 범위로 제한하고, 그 범위 안의 다양한 배경들을 종합적으로 고려하여 디자인 된 군용 위장패턴을 통해 간접적으로 고준위 특성 간 거리를 계산 및 최소화할 수 있다. 이러한 위장 적대적 패치 최적화는 배경과 가깝게 최적화하는 방법론에 따라 크게 두 방법으로 분류된다. 첫 번째 방법은 배경과 적대적 패치 사이의 고준위 특성 간 거리를 직접적으로 최소화하는 것이다. Fig. 2 (a)의 배경 중심 위장형 적대적 최적화 경로가 무작위 패턴으로부터 배경과의 거리는 최소화하고, 물체와의 거리는 최대화하는 최적화 방법론을 표현하고

있다. 두 번째 방법은 군용 위장패턴과 적대적 패치 사이의 시각적 거리를 직접적으로 최소화하는 것이다. 이는 군용 위장패턴이 다양한 설문과 통계적 실험을 통해 배경과 높은 시각적 유사성을 가지도록 디자인 되었기에 적용될 수 있는 최적화 방법이다. Fig. 2 (a)의 군용 패턴 중심 위장형 적대적 최적화 경로를 통해 무작위 패턴으로부터 배경과 고준위 특성 간 거리가 상당히 작은 군용 위장패턴들 중 한 위장패턴과의 거리를 최소화하고, 물체와의 거리는 최대화하는 과정을 표현하였다. 두 종류의 위장 적대적 패치 최적화 과정들은 2.2장에서 제시된 군용 위장패턴 성능평가 지표 기반의 손실함수들을 사용하며, 이를 통해 위장을 고려하지 않은 적대적 패치 최적화 과정과 달리 사람의 인지와 인공신경망의 탐지를 모두 무력화할 수 있는 적대적 패치 최적화를 기대할 수 있다. 첫 번째 방법론에 따라 Fig. 5 (a)의 CamAdv1~4가, 두 번째 방법론에 따라 Fig. 5 (b)의 CamAdv5~8이 생성되었다.

3. 객체탐지 모델 공격 실험

본 장에서는 앞서 기술한 적대적 공격 방법론을 바탕으로 효율적인 적대적 패치 최적화를 수행하여 전차 단일 클래스에 대한 객체탐지 모델에 대한 공격 성능을 확인하였고, 위장패턴 및 위장성능 평가 지표들을 각기 다른 방법으로 적용하여 적대적 공격을 위한 적대적 패치와 위장 기법이 적용되지 않은 일반적인 적대적 패치를 생성하고, 그 공격 성능 변화에 대해 기술하는 실험을 수행하였다. 학습용 데이터와 공격 대상 모델은 4장에서 기술한 바와 동일하다. 모든 연산은 2080 Ti GPU 4개가 장착된 단일 워크스테이션에서 수행하였다.

3.1 데이터셋 및 모델

전차를 탐지하는 객체탐지 모델은 Yolo v5s[17]를 사용하였다. Yolo v5는 가장 최신의 객체탐지 모델 중 하나이며, 사용되는 모델의 깊이에 따라 small, medium, large 등으로 구분되며 본 연구에서는 학습데이터의 규모를 고려하여 small 모델을 사용하였다. 해당 모델을 COCO dataset으로 일반적인 객체탐지 기능을 위한 사전학습을 수행하고, 일반적으로 공개된 전차 이미지들을 인터넷 상에서 수집하고, 전차 형상이 충분히 드러난 900장의 전차 이미지를 학습용 600장, 검증용 300장으로 분리하여 전차 탐지 성능을 극대화하도록 객체탐지 모델의 파라미터 값을 조정하였다. 적대적 패치는 공격 대상이 되는 물체 위에 부착하는 공격방식이므로 이러한 공격 방식을 디지털 상의 이미지에 모사하여, 객체의 위치를 주석(Annotation) 정보를 바탕으로 확인하고 그 중앙에 사각형태의 적대적 패치를 부착하는 공격방법[15]을 적용하였다. 이 때 적대적 패치를 실물 적용하는 과정에서의 관측 거리 및 방향, 외부 조명의 영향을 포괄하기 위해 적대적 패치의 크기, 명도, 부착 각도에 일정 범위 내의 무작위성이 부여되어 강건한 적대적 패치 최적화를 수행하였고 객체탐지 성능 저하, 적대적 패치 실물 적용, 그리고 위장성능 평가 지표를 기반으로 한 손실함수들을 정의하고 그 합을 최소화하도록 각 픽셀들의 값을 최적화했다. 적대적 패치의 해상도에 따른 공격 성능의 실험 결과는 Table 1에서 표시하였고 가장 높은 공격 성능을 보인 300×300의 해상도를 모든 실험에서 채택하였다.

Table 1. Object detection results by patch resolution

Resolution	AP50 [%]	AP [%]
50×50	86.4	54.9
100×100	31.5	16.1
150×150	24.3	13.1
200×200	24.3	12.2
250×250	23.5	12.1
300×300	**19.8**	10.5
350×350	23.2	11.6
400×400	19.9	10.2

3.2 적대적 패치 최적화

전차 단일 클래스에 대한 객체탐지 모델로 학습된 Yolov5s 모델에 대하여 위장을 고려하지 않는 적대적 패치 최적화를 수행하였다. 총 1000회(Epoch)의 최적화 과정을 반복하여 패치가 형성되었으며, 각 반복마다 생성된 적대적 패치들 중 775 epoch의 적대적 패치가 기존 AP50 기준 94.8 %의 높은 탐지 성능을 보이던 객체탐지 모델의 성능을 AP50 19.8 %로 크게 감소시켜 가장 공격 성능이 높았다. 해당 패치가 적대적 패치가 전차 중앙에 부착된 이미지들을 Fig. 3에서 확인할 수 있다.

Fig. 3. Visualization of normal adversarial patch

(a) epoch 10 (b) epoch 100 (c) epoch 1000

Fig. 4. Adversarial patch by epoch

100 epoch 단계에서 생성된 적대적 패치의 경우에도 AP50 기준으로 20.1 %로 객체탐지 성능을 하락시켜, 100 epoch 내로 적대적 패치의 대부분의 공격 성능 향상은 100 epoch 이내에 이루어지며, 그 뒤로 반복되는 최적화 과정에서는 형태와 공격 성능은 큰 변동이 없었지만 실물 적용 가능성을 극대화하는 손실함수들 L_{nps}과 L_{tv}이 감소하며 보다 부드러운 색상 분포와 연속적인 구조를 가지는 적대적 패치가 형성되는 것 확인하였으며 그 형상들을 Fig. 4에서 비교할 수 있다.

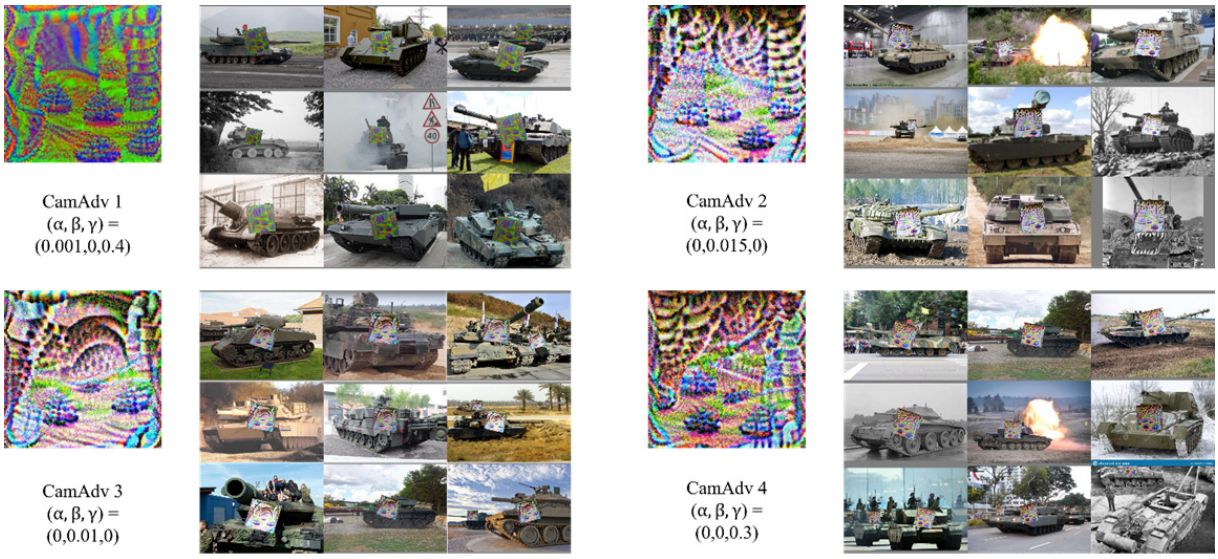
3.3 적대적 위장형 패치 최적화

전차 단일 클래스에 대한 객체탐지 모델로 학습된 Yolov5s 모델을 대상으로 하는 적대적 공격을 수행하기 위한 위장형 적대적 패치 최적화를 Fig. 2의 배경 중심, 군용 패턴 중심 위장형 적대적 최적화 경로에 따라 수행하였다. Fig. 5에 최적화 경로 및 $L_{camouflage}$의 세부 손실함수들에 대한 가중치 α, β, γ에 따라 서로 다르게 최적화된 적대적 패치의 형상과 그 패치가 전차에 부착되어 적용된 사진들을 확인할 수 있다. Fig. 5의 각 패치들은 위장을 고려하지 않은 적대적 패치 최적화 실험에서 100 epoch 내에 공격성능이 충분히 극대화 된다는 것을 확인하였기에 최적화 과정의 반복은 100 epoch으로 한정하고 공격성능을 비교하였다. 각 군용 위장패턴에 맞춰 최적화된 CamAdv 5-8의 형태가 군용 위장패턴과 상당히 유사함을 확인할 수 있다.

위장 적대적 최적화로 생성된 8종류의 적대적 패치들(CamAdv 1-8)의 성능을 Table 2에 표기하였다. CamAdv 1-4은 위장 적대적 최적화 I에 따라, CamAdv 5-8은 군용 패턴 중심 위장형 적대적 최적화에 따라 최적화된 패치들이다. 전반적인 적대적 공격 성능을 위장을 고려하지 않은 적대적 최적화를 통해 얻어진 Normal 패치와 비교해보면, Normal 패치의 공격 성능보다는 다소 낮지만 객체탐지 모델의 성능을 상당부분 감소시키는 것을 확인할 수 있다. 특히 CamAdv 2와 CamAdv 6의 경우에는 Normal 패치와 비교해도 각각 1.9 %, 2.1 %만의 성능 차이를 보여 두 종류의 최적화 방법론에서 최고의 공격 성능을 보였다. 또한 위장 적대적 최적화를 거친 각 패치들이 높은 공격 성능을 유지하면서도 위장을 고려하지 않은 적대적 최적화를 거친 Normal 패치에 비해 사람의 인지에도 상대적으로 잘 탐지되지 않는 위장성능을 갖추었음을

Table 2. Object detection results of various camouflaged adversarial patches

Patch	AP50 [%]	AP [%]
None	94.8	69.5
Normal	20.1(-74.7)	10.4(-59.1)
CamAdv 1	25.4(-71.4)	13.7(-55.8)
CamAdv 2	**22.0(-72.8)**	11.5(-58.0)
CamAdv 3	22.9(-71.9)	11.6(-57.9)
CamAdv 4	22.5(72.3)	11.6-57.9)
CamAdv 5	32.5(-62.3)	17.3(-52.2)
CamAdv 6	**23.0(-71.8)**	12.9(-56.6)
CamAdv 7	26.9(-67.9)	15.5(-54.0)
CamAdv 8	23.6(-71.2)	15.9(-53.6)



(a) Background centric optimization

(b) Military pattern centric optimization

Fig. 5. Visualization of various camouflaged adversarial patches

확인하였다. 기존의 Normal 패치의 경우에는 공격 대상이 되는 전차가 녹색 및 청색 계열의 색상 분포를 가지는 경우가 많기에 반대로 적색 계열의 특징이 강조되도록 최적화됨을 확인할 수 있다. 적색 픽셀이 많을수록 기존의 객체탐지 모델이 전차가 아니라고 판단하기 쉽기 때문이며, 이로 인해 전차라는 물체 뿐 아니라 배경과도 상이한 특성을 갖게 된다. 반면 CamAdv 1-4의 경우 배경과의 고준위 특성 간 거리를 최소화하는 배경 중심 위장형 적대적 최적화에 따라 생성되었기 때문에 Normal 패치에 비해 적색 픽셀이

줄어들고 녹색 및 청색이 증가하여 위장성능을 다소 개선시켰음을 확인하였다. 또한 CamAdv 5-8의 경우 전차가 주로 있는 숲 등의 배경과 유사하도록 설계된 군용 위장 패턴과의 고준위 특성 간 거리를 최소화하는 군용 패턴 중심 위장형 적대적 최적화를 따라 최적화되었고, 기존의 군용 위장 패턴과 상당히 유사한 형태를 띠면서도 공격 성능을 상당부분 유지하였다.

위장을 고려하지 않은 적대적 최적화와 위장형 적대적 최적화에 의해 생성된 적대적 패치를 종합적으로 비교해 보면 적대적 공격 성능은 일반적인 적대적 최적화에 따라 생성된 적대적 패치가 다소 높으나 배경 중심 위장형 적대적 최적화 경로와 군용 패턴 중심 위장형 적대적 최적화 경로에 따라 생성되는 적대적 패치들도 손실함수의 계수들을 적절히 조절하면 거의 유사한 수준의 공격 성능을 낼 수 있음을 확인할 수 있다. 위장성능은 위장을 고려하지 않은 적대적 최적화에 따른 비현실적 특성이 극대화되는 단점을 두 종류의 위장형 적대적 최적화들에서는 모두 상대적으로 개선되었으며 특히 군용 패턴 중심 위장형 적대적 최적화에 따라 군용 위장 패턴 형태로 최적화되는 적대적 패치의 위장성능이 가장 높았다.

본 연구에서 활용한 데이터셋의 크기가 제한적이기에 배경 중심 위장형 최적화의 효율성이 제한되어 위장성능 개선 효과가 상대적으로 적었지만 배경에 대해 더욱 높은 다양성과 범용성을 가지는 대규모 데이터셋에 대해서는 보다 효율적으로 적용될 수 있을 것으로 기대한다. 반면 군용 패턴 중심 위장형 적대적 최적화에 따른 최적화는 전차와 같이 주로 관측되는 배경의 특성이 일관되고, 해당 배경과 유사하도록 전문적인 연구를 통해 생성된 위장패턴이 사전에 존재해야 한다는 한계점을 가지지만 적대적 패치가 배경과 유사한 특징들을 가지도록 유도할 수 있는 위장패턴이 없는 경우에도 적용할 수 있다는 장점을 가짐을 확인하였다. 공격 대상이 되는 물체의 종류, 최적화에 사용될 데이터셋의 규모 등을 종합적으로 고려하여 본 논문에서 제시한 두 종류의 위장형 적대적 최적화 방법론들을 복합적으로 사용하여야 할 것이다.

4. 결 론

본 연구에서는 적대적 공격기술과 국방분야의 위장성능 평가기술에 대한 기술분석을 수행하였고 전차를 잘 탐지하도록 학습된 객체탐지 모델에 대하여 효과적인 적대적 공격을 수행할 수 있는 위장을 고려하지 않은 적대적 최적화 방법론과 사람의 인지특성을 함께 고려하는 위장형 적대적 최적화 방법론에 대한 유효성을 평가를 수행하였다. 또한 제시한 방법론들을 실제로 구현할 수 있도록 군용 위장패턴의 위장성능 평가 지표들을 기반으로 한 손실함수들을 통해 고준위 특성 간 거리를 간접적으로 측정하고, 각 방법론들을 기반으로 적대적 패치를 생성하는 실험을 수행하였다. 본 논문에서 제시한 적대적 위장형 패치들은 기존 적대적 패치의 공격 성능(75 %)과 유사한 수준의 공격 성능(최대 73 %)을 유지하면서도 위장 성능을 높였으며 기존에 사용되어 온 적대적 패치를 통한 공격 방법이 본래 목적에 반하여 육안에 의한 색적에 취약해지는 한계점을 극복할 수 있음을 실험적으로 증명하였다. 최적화 과정을 통해 생성된 적대적 패치와 최적생성 알고리즘들은 국방분야 객체탐지 딥러닝 모델에 대한 잠재적인 적대공격에 대한 강건성 검증에 활용이 가능할 것으로 판단된다. 또한 군용 위장패턴 기반의 적대적 패치 생성기술을 활용하여 특정조건에서 요구되는 적대적 위장패치를 효과적으로 생성할 수 있을 것으로 기대된다.

References

[1] Rey Reza Wiyatno et al., “Adversarial Examples in Modern Machine Learning: A Review,” arXiv preprint arXiv:1911.05268, 2019.

[2] Eykholt, Kevin, et al., “Robust Physical-World Attacks on Deep Learning Visual Classification,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[3] Den Hollander, Richard, et al., “Adversarial Patch Camouflage Against Aerial Detection,” Artificial Intelligence and Machine Learning in Defense Applications II, Vol. 11543, International Society for Optics and Photonics, 2020.

[4] Zhang, Yang, et al., “CAMOU: Learning Physical Vehicle Camouflages to Adversarially Attack Detectors in the Wild,” International Conference on Learning Representations, 2018.

[5] Wu, Tong, et al., “Physical Adversarial Attack on

Vehicle Detector in the Carla Simulator," arXiv preprint arXiv:2007.16118, 2020.

[6] Duan, Ranjie, et al., "Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[7] Lovisotto, Giulio, et al., "SLAP: Improving Physical Adversarial Examples with {Short-Lived} Adversarial Perturbations," 30th USENIX Security Symposium (USENIX Security 21). 2021.

[8] Sayles, Athena, et al., "Invisible Perturbations: Physical Adversarial Examples Exploiting the Rolling Shutter Effect," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[9] Tu, James, et al., "Physically Realizable Adversarial Examples for Lidar Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[10] Mariani, Giorgio, et al., "Generating Adversarial Surfaces via Band-Limited Perturbations," Computer Graphics Forum, Vol. 39, No. 5, 2020.

[11] Liu, Hsueh-Ti Derek, et al., "Adversarial Geometry and Lighting Using a Differentiable Renderer," CoRR, abs/1808.02651, 2018.

[12] Alexander Toet and Maarten A Hogervorst, "Review of Camouflage Assessment Techniques," In Target and Background Signatures VI, Volume 11536, page 1153604. International Society for Optics and Photonics, 2020.

[13] Chiuhsiang Joe Lin, Chi-Chan Chang, and Yung-Hui Lee, "Evaluating Camouflage Design Using Eye Movement Data," Applied Ergonomics, 45(3):714-723, 2014.

[14] Chi-Chan Chang, Yung-Hui Lee, Chiuhsiang Joe Lin, Bor-Shong Liu, and Yuh-Chuan Shih, "Visual Assessment of Camouflaged Targets with Different Background Similarities," Perceptual and Motor Skills, 114(2):527-541, 2012.

[15] Simen Thys, Wiebe Van Ranst, and Toon Goedem, "Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0-0, 2019.

[16] Kim, Jeonghun, et al., "Camouflaged Adversarial Attack on Object Detector," 2021 21st International Conference on Control, Automation and Systems (ICCAS). IEEE, 2021.

[17] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, October 2020.