Journal of the KIMST, Vol. 24, No. 6, pp. 591-601, 2021
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2021.24.6.591

Research Paper 센서·신호처리 부문

딥러닝 기반 적외선 객체 검출을 위한 적대적 공격 기술 연구

김호성*,1) · 현재국1) · 유현정1) · 김춘호1) · 전현호2)

1) 국방과학연구소 미사일연구원
2) 한국항공우주연구원 위성우주탐사체계설계부

Adversarial Attacks for Deep Learning-Based Infrared Object Detection

Hoseong Kim*,1) · Jaeguk Hyun1) · Hyunjung Yoo1) · Chunho Kim1) · Hyunho Jeon2)

1) *The 1st Research and Development Institute, Agency for Defense Development, Korea*
2) *Satellite & Space Exploration Systems Engineering and Architecture R&D Division, Korea Aerospace Research Institute, Korea*

(Received 23 June 2021 / Revised 13 September 2021 / Accepted 19 November 2021)

Abstract

Recently, infrared object detection(IOD) has been extensively studied due to the rapid growth of deep neural networks(DNN). Adversarial attacks using imperceptible perturbation can dramatically deteriorate the performance of DNN. However, most adversarial attack works are focused on visible image recognition(VIR), and there are few methods for IOD. We propose deep learning-based adversarial attacks for IOD by expanding several state-of-the-art adversarial attacks for VIR. We effectively validate our claim through comprehensive experiments on two challenging IOD datasets, including FLIR and MSOD.

Key Words : Adversarial Attack(적대적 공격), Infrared Object Detection(적외선 객체 검출), Deep Learning(딥러닝)

1. 서 론

지난 수 년간 심층 신경망(Deep Neural Network)은 급격한 기술 발전을 바탕으로 영상 분류, 객체 검출 등 다양한 분야에서 널리 활용되고 있다[1]. 하지만 최근에는 Fig. 1의 (b)와 같이 육안으로 식별할 수 없는 수준의 작은 퍼터베이션(perturbation) 만으로도 심층 신경망의 성능을 크게 감소시킬 수 있는 적대적 공격(Adversarial Attack)에 관한 연구 역시 활발히 진행되고 있다[2,3]. 또한 적대적 공격은 일반적인 가우시안 잡음(Gaussian noise)에 비해 효과적이며 심층 신경망에 특화되어 있다는 것이 실험적으로 증명되었다[3]. 이는 딥러닝 기반 객체 검출 연구의 취약성을 단적으로 제시하며, 심층 신경망을 사용하는 경우에는 적대적 공격 가능성에 대해 반드시 염두에 두어야 한다.

적대적 공격 연구는 크게 영상 분류와 객체 검출을 위한 방법으로 나뉜다. 영상 분류는 이미지 내 단일 객체의 클래스 정보만을 예측하지만 객체 검출은 다수 객체의 클래스 정보와 위치를 동시에 예측해야 하

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

(a) Clean image (b) Imperceptible perturbation (c) Adversarial image

Fig. 1. An adversarial example by adding an imperceptible perturbation for infrared object detection

므로 훨씬 도전적인 연구 분야이다. 그런데 객체 검출을 위한 적대적 공격은 심층 신경망의 모델별로 공격 방법을 특정지어야 하므로 호환성(compatibility)이 현격히 떨어지는 문제점을 갖는다. 예를 들어 Faster R-CNN[4]과 같은 two-stage 검출기를 위한 공격 방법은 one-stage 검출기에 적용하기 어렵다. 이와 반대로 영상 분류를 위한 적대적 공격은 심층 신경망의 종류에 관계없이 활용 가능하므로 호환성이 높다. 즉, 하나의 기술로 다양한 모델에 적용 가능하다.

현실적으로 수많은 네트워크에 대해 각각 다른 방법을 적용해야하는 기존의 객체 검출용 적대적 공격 기술은 다양한 분야에 적용하기 어렵다. 따라서 본 논문에서는 호환성을 고려하여 네트워크의 종류에 상관없이 적용 가능한 영상 분류용 적대적 공격 기술을 확장하여 객체 검출에 활용하였으며, 다양한 분야에서 널리 활용되는 적외선 도메인을 사용하였다.

본 연구의 contribution은 다음과 같다. (1) 딥러닝 기반 적외선 객체 검출의 취약점을 분석하였고, (2) 호환성 높은 영상 분류용 적대적 공격 기술을 확장하여 객체 검출에 활용하였으며, (3) 다양한 정량적·정성적 실험을 통해 제안하는 방법의 우수성을 효과적으로 검증하였다는 것이다.

2. 관련 연구

본 장에서는 가시광 및 적외선 도메인을 위한 딥러닝 기반 객체 검출 분야에 대한 기본적인 내용과 관련 연구를 서술한다. 또한 영상 분류 및 객체 검출을 위한 적대적 공격 분야에 대한 핵심적인 내용과 기존 연구들에 대해 간략히 설명한다.

2.1 딥러닝 기반 객체 검출

객체 검출은 다수 객체에 대한 클래스와 위치 정보를 동시에 예측해야 하는 기술이다. 심층 신경망을 이용한 객체 검출 기술은 ImageNet[5], COCO[6]와 같이 양질의 대용량 데이터셋이 공개된 이후로 모델의 특징 표현(feature representation) 능력을 비약적으로 향상시킴으로써 복잡한 환경 속에서도 높은 수준의 정확도를 갖게 되었다. 본 논문에서는 최신 기술인 EfficientDet[7]과 YOLOv4[8]를 기본 검출기로 사용하였다. 일반적인 객체 검출 모델은 backbone, neck, head로 구성되어 있다. 입력 영상으로부터 backbone을 통해 특징을 추출하고 neck은 backbone과 head 사이의 연결 고리 역할을 함과 동시에 서로 다른 레이어 사이의 관계를 고려하여 특징을 혼합함으로써 객체 검출 성능을 높이는 데 매우 중요한 역할을 한다. 마지막으로 head 단계에서는 추출된 특징을 종합하여 다양한 크기의 객체를 검출하는 과정을 수행한다.

한편 대부분의 객체 검출 연구가 가시광 도메인에 한정되어 있는 반면 본 논문에서는 다양한 분야의 응용에 초점을 맞추어 적외선 도메인을 사용하였다. 적외선 객체 검출 연구는 단순히 가시광 객체 검출 기술을 적외선 도메인에 도입하여 활용하는 수준에 머물러 있다. 대표적으로 멀티 스펙트럴 객체 검출(Multispectral Object Detection, MSOD)[9], 심층 멀티모달 객체 검출(Deep Multi-modal Object Detection)[10], FLIR(Forward-Looking Infrared)[11] 등이 있다. 적외선 객체 검출 연구가 어려운 이유는 적외선 도메인의 대용량 데이터셋 부재로 인한 사전학습(pre-training)이 어렵다는 점이다. 이를 위한 해결 방안으로 가시광 도메인의 대용량 데이터셋인 ImageNet과 COCO를 활용한 전이학습(transfer learning)을 수행하였다.

2.2 적대적 공격

적대적 공격은 눈으로 인지할 수 없는 수준의 퍼터베이션을 입력 영상에 추가하여 심층 신경망의 정확도를 약화시키는 기술이다. 적대적 공격의 중요한 특징은 일반적으로 공격 대상을 의미하는 위협 모델(threat model)별로 각각 학습시켜야 한다는 것이다. 예를 들어 EfficientDet을 공격하여 생성한 적대적 영상(adversarial image)은 EfficientDet의 성능에 심각한 타격을 주지만 YOLOv4의 성능에는 큰 영향을 주지 못한다.

적대적 공격 연구의 큰 흐름은 영상 분류와 객체 검출 분야로 나뉘며 대부분은 영상 분류용 적대적 공격 연구에 집중되어 있다[2,12-20]. 예를 들어, FGSM[2](i), DDN[12](ii), BIM[13](iii), PGD[14](iv), Newton[15]은 경사 하강법(gradient descent), Inversion[16]은 픽셀값 역변환, VAT[17](v)는 쿨백-라이블러 발산(Kullback-Leibler divergence, D_{KL}), DeepFool[18]은 최근접 경계(nearest boundary)를 사용하여 적대적 공격을 수행하였다. 영상 분류용 공격은 단일 객체에 대한 클래스를 예측하는 단순한 구성으로 이루어져 있으므로 단일 공격 기술로 여러 심층 신경망 모델에 적용 가능한 장점을 갖는다.

한편 객체 검출을 위한 적대적 공격 연구에는 대표적으로 DAG(vi)[21], RAP[22](vii) 등이 있으며, 대부분 Faster R-CNN과 같이 영역 제안 네트워크(region proposal network, RPN) 방식의 two-stage 검출기를 위한 공격 모델이다. 하지만 이러한 기존 공격 방법들은 호환성이 낮은 단점이 있으며, 최근 가장 좋은 성능을 보여주는 one-stage 검출기인 EfficientDet, YOLOv4 등에 적용할 수 없다는 한계가 있다. 물론 DPatch[23]와 같은 one-stage 검출기를 위한 공격 모델에 관한 연구도 있지만 이는 기존 연구인 디지털 공격과 달리 물리적 공격으로써 영상에 실제 패치 형태를 추가함으로써 원본 영상과 육안으로도 큰 차이를 보이는 단점을 갖는다. 따라서 본 논문에서는 호환성이 높은 영상 분류용 적대적 공격 기술을 확장하여 one-stage 객체 검출에 활용하는 방안을 제안한다†.

(i) FGSM[2]: Fast Gradient Sign Method
(ii) DDN[12]: Decoupled Direction and Norm
(iii) BIM[13]: Basic Iterative Method
(iv) PGD[14]: Progected Gradient Descent
(v) VAT[17]: Virtual Adversarial Training
(vi) DAG[21]: Dense Adversary Generation
(vii) RAP[22]: Robust Adversarial Perturbation
†RPN을 고려하면 two-stage 검출기에도 적용 가능함

3. 적외선 객체 검출을 위한 적대적 공격

본 장에서는 적외선 객체 검출 및 영상 분류용 적대적 공격에 관한 내용과 제안하는 적외선 객체 검출용 적대적 공격 기술 및 평가 지표에 대해 설명한다.

3.1 적외선 객체 검출

적외선 객체 검출 연구는 대용량 데이터셋 부재로 인해 심층 신경망 모델의 성능 향상에 필수 요소인 사전학습을 적용하기 어렵다. 이를 해결하기 위해 가시광 대용량 데이터셋인 ImageNet과 COCO를 활용한 전이학습을 수행하였다. 구체적으로 ImageNet, COCO 데이터셋에서 사전학습한 모델을 각 적외선 데이터셋에서 fine-tuning 하였다. 실험에 사용한 검출기는 최신 기술인 EfficientDet[7]과 YOLOv4[8]를 사용하였다. EfficientDet을 기준으로 FLIR 데이터셋에서 전이학습 유무에 따른 성능 변화 실험을 진행한 결과 평균 정밀도(Average Precision, AP)와 AP_{50}에 대해 전이학습을 사용하였을 경우 각각 12.5 %, 17.7 % 성능이 향상되었다.

3.2 적외선 영상 분류를 위한 적대적 공격

적대적 공격은 일반적인 가우시안 잡음에 비해 훨씬 더 작은 퍼터베이션만으로도 심층 신경망의 정확도에 막대한 피해를 입힐 수 있는 기술이다[3]. 중요한 특징은 공격 대상을 뜻하는 위협 모델이 여러 개일 경우, 각 위협 모델별로 학습시켜야 한다는 것이다. 본 연구에서는 호환성이 높은 영상 분류용 적대적 공격 방법을 확장하여 객체 검출에 활용하는 방안을 제안한다.

기본적으로 적대적 공격 기술은 심층 신경망 F와 입력 영상 $x \in [0,1]^n$, 클래스 정보 y에 대해 적대적 영상 $\tilde{x} \in [0,1]^n$을 생성하며, 모델 손실 함수 J, 퍼터베이션 δ에 대해 다음의 조건을 만족해야 한다.

$$\max_{x+\delta \in B(x,\epsilon)} J(F(x+\delta), y). \quad (1)$$

조건 (1)의 ϵ은 인지할 수 없는 충분히 작은 값, $B(x,\epsilon)$은 x 주위의 ϵ-ball을 의미한다. 즉, 입력 영상에 아주 작은 변화 δ를 주어 모델을 속일 수 있는 적대적 영상 $\tilde{x}$를 생성하는 것이 핵심이다. 보통 $B(x,\epsilon)$의 거리 단위는 L_0, L_2, L_∞를 사용하며, J는 영상 분류를 위한 손실 함수를 활용한다.

대표적인 영상 분류용 적대적 공격 기술에는 크게 경사 하강법[2,12-15], 픽셀값 역변환[16], 쿨백-라이블러 발산[17], 최근접 경계[18], 크기 재조정(rescaling)[19] 등이 있으며 각각에 대한 설명은 다음과 같다.

FGSM[2]은 손실 함수의 경사(gradient) $\nabla_x J$를 이용하여 $J(F(x),y)$를 최대화시키며, 생성된 $\tilde{x}$는 식 (2)와 같으며, 거리 단위는 L_∞를 사용한다.

$$\tilde{x} = x + \epsilon sign(\nabla_x J(F(x),y)). \tag{2}$$

DDN[12]은 매 단계 k마다 손실 함수의 경사 $\nabla_{\tilde{x}_k} J$에 따라 δ_k, $\tilde{x}_k$를 조정하며, $\tilde{x}_k$의 적대적 여부에 따라 ϵ_k을 증감시킨다. 구체적으로, $\tilde{x}_k$가 적대적인 경우 최적의 $\tilde{x}_{k+1}$를 구하기 위해 ϵ_{k+1}을 감소시키며, 반대로 $\tilde{x}_k$가 적대적이지 않은 경우 적대적인 $\tilde{x}_{k+1}$를 생성하기 위해 ϵ_{k+1}을 증가시킨다. DDN은 기존 방법에 비해 학습 속도가 매우 빠르며 최종적으로 생성된 $\tilde{x}$는 원본 영상 x와 L_2-norm이 가장 작은 적대적 영상이다.

BIM[13]은 ϵ보다 작은 스텝 크기(step size) α에 대해 반복하는(iterative) FGSM을 수행하며 생성된 $\tilde{x}_{k+1}$은 식 (3)과 같으며, 거리 단위는 L_∞를 사용한다.

$$\tilde{x}_{k+1} = Clip_{(x,\epsilon)}\left\{\tilde{x}_k + \alpha sign(\nabla_{\tilde{x}_k} J(F(\tilde{x}_k),y))\right\}, \tag{3}$$

$$Clip_{(x,\epsilon)}\{\tilde{x}_k\} = \min(\max(\tilde{x}_k, x-\epsilon), x+\epsilon). \tag{4}$$

식 (4)에서 $Clip$ 함수는 $\tilde{x}_k$의 픽셀값들이 원본 영상 x 주위의 ϵ-ball 내에 있도록 잘라내는 역할을 한다.

PGD[14]는 FGSM과 달리 손실 함수의 경사를 여러번 사용하여 매우 효과적인 적대적 공격을 수행한다. 우선 입력 영상에 균등 분포 U로부터 임의의 잡음을 추가하여 초기 영상 $\tilde{x}_0 = x + U(-\epsilon,\epsilon)$를 생성하고, $k+1$ 단계에 대한 적대적 영상 $\tilde{x}_{k+1}$은 다음과 같다.

$$\tilde{x}_{k+1} = \prod_{B(x,\epsilon)} \left\{\tilde{x}_k + \alpha sign(\nabla_{\tilde{x}_k} J(F(\tilde{x}_k),y))\right\}. \tag{5}$$

식 (5)에서 α는 스텝 크기, $\prod_{B(x,\epsilon)}$는 $B(x,\epsilon)$에 투영시킨 것을 의미하며, 거리 단위는 L_∞를 사용한다.

Newton[15]은 선형 근사(linear approximation)와 경사 하강법을 이용하여 반복적으로 δ_k, $\tilde{x}_k$를 조정하며, 거리 단위는 L_2를 사용한다. 다른 방법들과 달리 클래스 정보 y에 대한 심층 신경망의 softmax 값이 매 단계 k마다 점점 작아지도록 학습한다.

Inversion[16]은 단순히 픽셀값 역변환을 이용하여 적대적 영상 $\tilde{x} = 1 - x$를 생성한다.

VAT[17]는 입력 영상 x와 적대적 영상 $\tilde{x}$에 대한 심층 신경망 예측값 $F(x)$, $F(\tilde{x})$ 사이의 쿨백-라이블러 발산 $D_{KL}(F(x)\|F(\tilde{x}))$ 값이 최대가 되며 모델을 속일 수 있는 $x+\delta \in B(x,\epsilon)$을 이용하여 적대적 영상 $\tilde{x} = x + \delta$를 생성한다. 거리 단위는 L_2를 사용한다.

DeepFool[18]은 최근접 경계와 선형 근사를 이용하여 적대적 영상을 반복적으로 조정하며, 거리 단위는 L_2를 사용한다.

$$\tilde{F}_c = F(x)_c - F(x)_y, \tag{6}$$

$$\tilde{w}_c = \nabla_x F(x)_c - \nabla_x F(x)_y. \tag{7}$$

식 (6), (7)에서 $F(x)_c$는 $c \in [1,2,...,C]$번째 클래스에 대한 모델 예측값을 의미하며, 최근접 경계의 초평면 $c^* = \arg\min_{c \neq y} |\tilde{F}_c| / \|\tilde{w}_c\|_2$에 대해 아주 작은 퍼터베이션 $\delta_k = (|\tilde{F}_{c^*}| / \|\tilde{w}_{c^*}\|_2^2)\tilde{w}_{c^*}$을 계산하고 매 단계 k마다 적대적 영상 $\tilde{x}_{k+1} = \tilde{x}_k + \delta_k$를 생성한다.

Rauber[19]는 크기 재조정과 $Clip$ 함수를 이용하여 식 (8)을 만족하는 $\eta \in R^n$를 구하고, 적대적 영상 $\tilde{x} = x + \eta\delta$를 생성한다. 거리 단위는 L_2를 사용한다.

$$\|Clip_{(x,\epsilon)}\{x+\eta\delta\} - x\|_2 = \epsilon. \tag{8}$$

식 (8)에서 δ는 가우시안 분포나 균등 분포에서 임의로 추출한 값을 사용하며, 각 방법에 따라 Rauber-G, Rauber-U로 나뉜다.

3.3 적외선 객체 검출을 위한 적대적 공격

객체 검출을 위한 적대적 공격 연구는 대부분 Faster R-CNN과 같이 영역 제안 방식의 two-stage 검출기에 한정되어있다[21,22]. 이러한 기존 방법들은 호환성이 낮아 다른 종류의 검출기에 적용할 수 없다. 따라서 본

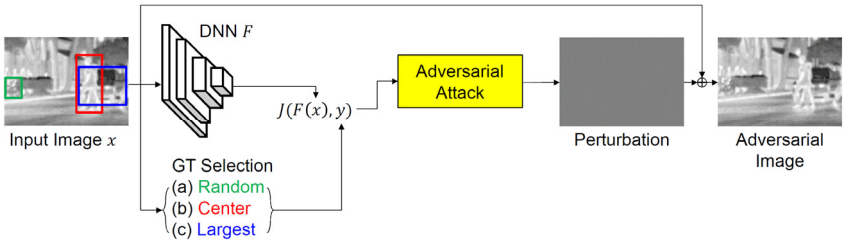


Fig. 2. The overview of our proposed adversarial attacks for infrared object detection

논문에서는 호환성이 높은 영상 분류용 적대적 공격 기술을 확장하여 one-stage 검출기에 적용 가능한 객체 검출용 적대적 공격 방법을 제안한다. 구체적으로, cross-entropy 손실 함수와 함께 영상 내에 존재하는 여러 개의 GT(Ground Truth) 바운딩 박스(bounding box) 중에서 (1) 임의(random) 선택, (2) 영상의 중심점과 가장 가까운 것(center) 선택, (3) 가장 크기가 큰 것(largest)을 선택하여 사용한다. 선택된 GT에 대해 손실 값을 계산하고 이를 최대가 되도록 하는 적대적 영상을 생성하며 이는 Fig. 2와 같다. 제안하는 GT 선택 방법을 통해 객체 검출 모델을 효과적으로 무력화시켰으며 폭넓은 정량적·정성적 실험을 통해 제안하는 방법의 우수성을 검증하였다. 한편 제안하는 방법은 기존 방법과 구분짓기 위해 기존 적대적 공격에 '+'를 더하여 표기한다. 예를 들어, FGSM+, DDN+ 등이 있다.

3.4 적대적 공격 평가 지표

일반적으로 적대적 공격 방법의 성능을 평가하는 지표로써 성능 하락폭과 퍼터베이션의 크기를 사용한다. 즉, 퍼터베이션의 크기가 작으면서 원본 영상 대비 적대적 영상의 성능 하락폭이 클수록 효율적인 적대적 공격 방법을 의미한다. 구체적으로 $\triangle_{AP} = AP_{\tilde{x}} - AP_x$는 두 영상 간 평균 정밀도(Average Precision, AP) 차이를 의미하며, $\|\delta\| = \|x - \tilde{x}\|_2 + 1e^{-10}$는 두 영상 간 퍼터베이션의 크기를 의미한다. $\|\delta\|$ 값이 작을수록 원본과 유사하고 $|\triangle_{AP}|$ 값이 클수록 정확도 감소량이 증가하므로 효과적인 적대적 공격 방법이라 할 수 있다.

4. 실 험

본 장에서는 실험에 사용한 데이터셋, 평가 지표, 구현 세부사항, 적대적 공격과 잡음 비교 실험, 적외선 객체 검출을 위한 적대적 공격, 적대적 공격 특성 분석, 원본 영상과 적대적 공격 영상 활성화 맵 비교, 적대적 공격 예시에 대해 설명한다.

4.1 데이터셋, 평가 지표, 구현 세부사항

본 논문에서는 딥러닝 기반 적외선 객체 검출의 취약성을 분석하기 위해 FLIR[11], MSOD[9] 데이터셋을 활용한다. FLIR 데이터셋은 3개의 클래스(person, bicycle, car)와 10,228개의 이미지, 79,300개의 바운딩 박스로 구성되어 있으며, 영상의 해상도는 640×512이다. MSOD 데이터셋은 멀티 스펙트럴 파장대역의 정보를 포함하고 있으며 본 실험에서는 FIR(Far Infrared)을 사용하였다. 구체적으로, 5개 클래스(person, bicycle, car, color_cone, car_stop)와 6,319개의 이미지, 28,271개의 바운딩 박스를 포함하며, 영상의 해상도는 640×480이다. 각 데이터셋에 대한 요약 정보는 Table 1과 같다.

Table 1. Details of FLIR and MSOD datasets

데이터셋	클래스 수	영상 수			바운딩 박스 수		
		학습	테스트	총합	학습	테스트	총합
FLIR[11]	3	8,862	1,366	10,228	67,618	11,682	79,300
MSOD[9]	5	4,964	1,355	6,319	21,560	6,711	28,271

본 실험을 위해 우선 적외선 객체 검출을 위한 심층 신경망(EfficientDet, YOLOv4)을 학습 데이터셋(FLIR, MSOD)에서 fine-tuning 한다. 그리고 학습된 모델과 테스트셋을 기반으로 각 적대적 공격을 수행하였을 때 평균 정밀도를 측정한다. 감소하는 평균 정밀도가 크고, 원본 영상과의 차이가 작을수록 효과적인 적대적 공격 방법이다. 또한 그 외에도 객체 검출의 주요 평가 지표인 COCO[6] 데이터셋의 평균 정밀도(Average Precision, AP), AP_{50}, AP_{75}, AP_S, AP_M , AP_L을 사용하였

다. 평균 정밀도 AP는 GT와 예측된 바운딩 박스 사이의 IoU(Intersection-over-Union) 값이 기준점(0.5부터 0.95까지 0.05씩 증가하여 총 10 단계) 이상인 경우에 대한 AP 값들의 평균으로 계산한다. AP_{50}, AP_{75}는 각각 IoU 기준점이 0.5, 0.75일 때의 AP 값을 의미한다.

AP_S, AP_M , AP_L은 픽셀 기준으로 물체 면적이 각각 $0\sim32^2$, $32^2\sim96^2$, 96^2 이상인 경우의 AP를 의미한다. 주요 평가 지표인 평균 정밀도 차이 크기 $|\triangle_{AP}|$는 클수록 좋으며, 퍼터베이션 크기인 $\|\delta\|$는 작을수록 좋다.

실험에 사용된 총 10개의 최신 적대적 공격 기술은 Inversion[16], FGSM[2], DDN[12], BIM[13], PGD[14], Newton[15], VAT[17], DeepFool[18], Rauber-G[19], Rauber-U[19]를 포함한다. 실험에 사용한 GPU는 1대의 NVIDIA Titan RTX이며, 적대적 공격의 한계점인 ϵ 값은 실험을 통해 결정하였으며 제안하는 Inversion+, FGSM+, DDN+, BIM+, PGD+, Newton+, VAT+, DeepFool+, Rauber-G+, Rauber-U+에 대해 FLIR 데이터셋에서는 각각 0.09, 0.004, 2.0, 0.008, 0.014, 1.0, 1.2, 0.01, 20.0, 30.0을 사용하였으며 MSOD 데이터셋에서는 각각 0.2, 0.0007, 2.0, 0.001, 0.0007, 0.4, 1.0, 0.003, 20.0, 20.0을 사용하였다. 객체 검출기는 최신 기술인 YOLOv4[8]와 EfficientDet-D2[7]를 사용하였다. 표기의 간결성을 위해 특별히 다른 언급이 없는 경우, EfficientDet-D2를 EfficientDet으로 표기한다. YOLOv4와 EfficientDet은 ImageNet과 COCO 데이터셋에서 pre-training한 모델을 사용하였다.

4.2 적대적 공격과 잡음 비교 실험

본 절에서는 적대적 공격과 일반적인 잡음 방법인 가우시안 잡음, 점잡음(salt and pepper noise)과의 공격 효과를 비교하였다. Table 2에서 EfficientDet 모델에 대한 가우시안 잡음과 점잡음의 $\triangle_{AP}$는 각각 -0.1 %, -0.3 %를 기록한 반면, 적대적 공격 방법인 FGSM+, DDN+, BIM+, PGD+에 대한 평균 $\triangle_{AP}$는 -4.0 %로써 잡음 방식에 비해 훨씬 큰 성능 하락폭을 기록하였다. 또한 퍼터베이션의 크기인 $\|\delta\|$ 역시 가우시안 잡음과 점잡음은 각각 3.03, 1.82를 기록한 반면, 적대적 공격 방법은 0.19만을 기록하였다.

정성적인 비교 실험 결과인 Fig. 3의 첫 번째 행은 원본 영상에 퍼터베이션을 추가한 결과 영상을 4배 확대한 것이며, 두 번째 행은 각 방법에 대한 퍼터베

Table 2. Performance comparison between baseline, noises, and adversarial attacks on FLIR dataset for EfficientDet

Methods	EfficientDet[7]		
	$\triangle_{AP}$	$\|\delta\|$	AP
Baseline (clean image)	0	$1e^{-10}$	35.8
Gaussian noise	-0.1	3.03	35.7
Salt and pepper noise	-0.3	1.82	35.5
Adversarial attacks	**-4.0**	**0.19**	**31.8**

(a) Baseline(clean image) (b) Gaussian noise (c) Salt and pepper noise (d) Adversarial attacks

Fig. 3. Examples of baseline, gaussian noise, salt and pepper noise, and adversarial attacks

이션 영상을 4배 확대한 것이다. Fig. 3의 (b) 가우시안 잡음과 (c) 점잡음은 식별 가능한 큰 퍼터베이션을 필요로 하는 반면에 적대적 공격 (d)는 육안으로 확인하기 어려운 수준의 퍼터베이션으로 공격 가능하다.

본 실험을 통해 잡음 방식은 육안으로 뚜렷하게 확인할 수 있을만큼 영상에 많은 변화를 가하지만 심층 신경망 기반 적외선 객체 검출기의 성능에는 큰 영향을 주지 못 한다는 것을 알 수 있다. 그러나 적대적 공격 방법은 육안으로 확인하기 어려운 아주 작은 변화를 가함에도 불구하고 모델의 성능에 큰 영향을 미치는 것을 확인하였다.

4.3 적외선 객체 검출을 위한 적대적 공격

3.3절에서 언급한 것과 같이 제안하는 방법은 (1) 임의 선택(random), (2) 중심점과 가장 가까운 것(center), (3) 가장 크기가 큰 것(largest)을 선택하여 적대적 공격을 수행한다. 이를 MSOD 데이터셋에서 baseline에 대해 비교 분석한 결과 Fig. 4와 같이 다양한 적대적 공격에 대해 기존 방법 대비 제안하는 세 가지 방법을 적용하였을 때 성능을 더욱 악화시키는 것을 확인하였다. 특히 Baseline은 평균적으로 -1.2 % 성능을 감소시킨 것에 비해 random, center, largest는 각각 -2.1 %, -2.0 %, -2.5 %로써 두 배에 가까운 성능 감소폭을 보였다. Largest의 성능이 가장 좋은 이유는 바운딩 박스의 크기가 클수록 영상 내에서 더 많은 정보를 갖고 있을 확률이 높고 결과적으로 모델의 특징 표현력에 가장 큰 영향을 주기 때문이다.

추가적으로, 본 절에서는 기존의 최신 영상 분류용 적대적 공격을 확장하여 적외선 객체 검출에 활용한 경우에 대해 제안하는 largest 방법을 기준으로 정량적으로 비교하였다. Table 3과 같이 FLIR 데이터셋에서 EfficientDet 기준으로 Inversion+는 성능 하락폭이 -5.9 %로 큰 편이지만 퍼터베이션 크기 $\|\delta\|$ 역시 42.8로 다른 방법들의 평균값인 0.17에 비해 250배 이상 크므로 상당히 비효율적인 방법이라 할 수 있다. 즉, Inversion은 공격 결과가 원본 영상과 눈에 띄게 구별되지만 다른 적대적 공격들은 육안으로 확인하기 어려운 수준으로 공격 가능하다. 다음으로 FGSM+, DDN+, BIM+, PGD+ 공격들은 각각 0.25, 0.01, 0.21, 0.32의 작은 퍼터베이션 크기만으로 -6.4 %, -1.4 %, -3.5 %, -4.5 %의 성능 감소를 보였다. 이는 육안으로 확인하기 어려운 매우 낮은 $\|\delta\|$로 $\triangle_{AP}$를 크게 하락 시키고, 공격 효율성을 극대화하였다고 할 수 있다. 한편 Newton+, VAT+, DeepFool+, Rauber-G+, Rauber-U+ 공격들은 각각 0.01, 0.02, 0.26, 0.10, 0.33의 퍼터베이션 크기로 -2.1 %, -4.1 %, -3.3 %, -1.1 %, -2.5 %의 성능 감소를 보였다. 결과적으로 원본과 거의 유사한 퍼터베이션 크기로 적외선 객체 검출을 위한 적대적 공격이 매우 효율적으로 작동한다는 것을 검증하였다.

마찬가지로 Table 4와 같이 MSOD 데이터셋에서도 제안하는 공격 방법이 매우 효과적이라는 것을 실험적으로 보였다. 우선 EfficientDet 기준으로 Inversion+는 성능 하락폭이 11.8 %로 매우 크지만 퍼터베이션

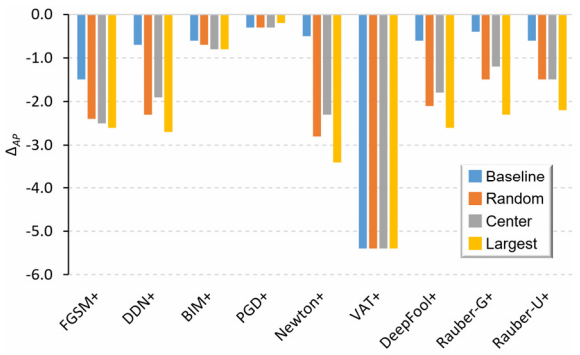


Fig. 4. Performance comparison with our proposed methods and baseline on MSOD dataset

Table 3. Adversarial attack performance w.r.t our largest GT box on FLIR dataset for EfficientDet

Methods		$\triangle_{AP}$	$\|\delta\|$	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Base	EfficientDet	0	$1e^{-10}$	35.8	65.7	35.9	19.4	44.7	62.5
Ours (Largest)	Inversion+	-5.9	42.8	29.9	55.6	26.7	14.0	38.6	61.0
	FGSM+	-6.4	0.25	29.4	56.2	26.7	16.1	37.4	57.2
	DDN+	-1.4	0.01	34.4	63.6	34.1	18.8	43.2	62.4
	BIM+	-3.5	0.21	32.3	60.2	30.6	16.6	41.4	57.9
	PGD+	-4.5	0.32	31.3	60.1	26.6	17.4	39.6	60.2
	Newton+	-2.1	0.01	33.7	62.4	32.5	19.1	42.2	62.4
	VAT+	-4.1	0.02	31.7	60.0	27.7	18.7	39.6	61.9
	DeepFool+	-3.3	0.26	32.5	60.9	31.0	17.0	41.3	61.0
	Rauber-G+	-1.1	0.10	34.7	64.1	34.5	19.1	43.4	62.6
	Rauber-U+	-2.5	0.33	33.3	61.5	32.5	18.7	41.3	62.1

Table 4. Adversarial attack performance w.r.t our largest GT box on MSOD dataset for EfficientDet

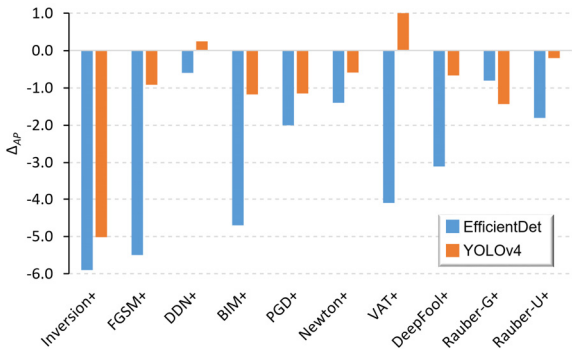
Methods		$\triangle_{AP}$	$\|\delta\|$	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Base	EfficientDet	0	$1e^{-10}$	40.4	70.5	40.5	11.3	42.2	63.9
Ours (Largest)	Inversion+	-11.8	92.5	28.6	51.4	28.3	5.3	27.7	54.6
	FGSM+	-2.6	0.01	37.8	67.3	37.4	11.0	40.1	58.8
	DDN+	-2.7	0.02	37.7	67.8	36.5	10.5	39.6	59.8
	BIM+	-0.8	0.01	39.6	69.4	39.7	11.4	41.4	62.3
	PGD+	-0.2	0.01	40.2	70.0	40.0	11.3	41.9	63.7
	Newton+	-3.4	0.01	37.0	66.6	35.9	10.9	39.5	57.7
	VAT+	-5.4	0.03	35.0	64.2	33.9	9.9	36.6	56.9
	DeepFool+	-2.6	0.04	37.8	67.9	37.3	11.0	40.6	58.7
	Rauber-G+	-2.3	0.24	38.1	67.6	37.6	10.8	39.9	61.0
	Rauber-U+	-2.2	0.24	38.2	67.8	37.5	10.9	39.8	61.2

크기 $\|\delta\|$가 92.5로 다른 방법의 평균값인 0.06에 1500배 이상 크므로 상당히 비효율적이라 할 수 있다. FGSM+, DDN+, BIM+, PGD+ 공격들은 모두 0.02 이하의 매우 작은 퍼터베이션 크기만으로 각각 -2.6 %, -2.7 %, -0.8 %, -0.2 %의 성능 감소를 보였다. 다음으로 Newton+, VAT+, DeepFool+, Rauber-G+, Rauber-U+ 공격들은 각각 0.01, 0.03, 0.04, 0.24, 0.24의 퍼터베이션 크기로 -3.4 %, -5.4 %, -2.6 %, -2.3 %, -2.2 %의 성능 감소를 보였다. 이는 원본과 거의 유사한 매우 작은 $\|\delta\|$로 $\triangle_{AP}$를 효과적으로 하락시키고, 공격 효율성을 높였다고 할 수 있다.

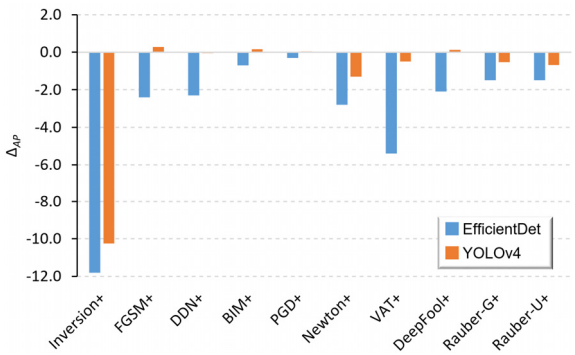
4.4 적대적 공격 특성 분석

2장에서 언급한 것처럼 적대적 공격의 주요 특징은 일반적으로 공격 대상을 의미하는 위협 모델별로 각각 학습해야 한다는 점이다. 이와 같은 특징이 적외선 객체 검출을 위한 적대적 공격에서도 발생하는지 검증하기 위해 EfficientDet을 공격하여 생성한 적대적 영상을 YOLOv4에 적용하여 성능을 비교하였다. FLIR, MSOD 데이터셋에 대한 실험 결과는 Fig. 5와 같다.

우선 Fig. 5의 (a) 결과를 보면 EfficientDet 모델 대비 YOLOv4 모델에서는 성능 하락폭이 크게 감소한 것을 알 수 있다. DDN+와 VAT+는 오히려 역효과를 가져왔



(a) FLIR



(b) MSOD

Fig. 5. Performance comparison with our proposed methods on FLIR and MSOD datasets for EfficientDet and YOLOv4(using adversarial images generated by attacking EfficientDet)

다. 그 이유는 앞서 언급한 것과 같이 적대적 공격은 공격 대상을 의미하는 위협 모델별로 학습하여 적용해야 하지만 본 실험에서는 EfficientDet을 공격한 결과를 YOLOv4에 적용하였기 때문에 공격 효과가 감소한 것이다. 이는 Fig. 5의 (b)와 같이 MSOD 데이터셋에서도 마찬가지로 EfficientDet 대비 YOLOv4에서는 성능 하락폭이 크게 감소하였다. 본 실험 결과를 통해 적대적 공격을 시도할 때는 심층 신경망 모델을 파악하고, 방어 시에는 모델을 은폐해야 한다는 것을 알 수 있다.

추가적으로 Fig. 6과 같이 각 적대적 공격의 임계값(threshold)에 따른 성능 감소폭 $\triangle_{AP}$ 변화 실험을 진행한 결과, 임계값이 커질수록 성능 감소폭도 같이 커지는 것을 확인하였으며 특히 FGSM+, BIM+, PGD+의 경우에는 임계값이 증가할수록 성능 감소폭이 -20 %에 가까울 정도로 매우 커지는 것을 확인할 수 있다.

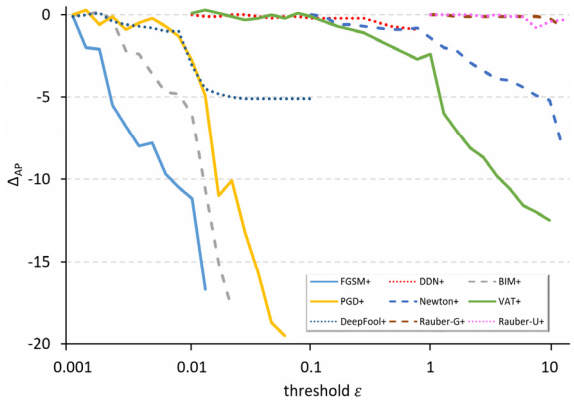
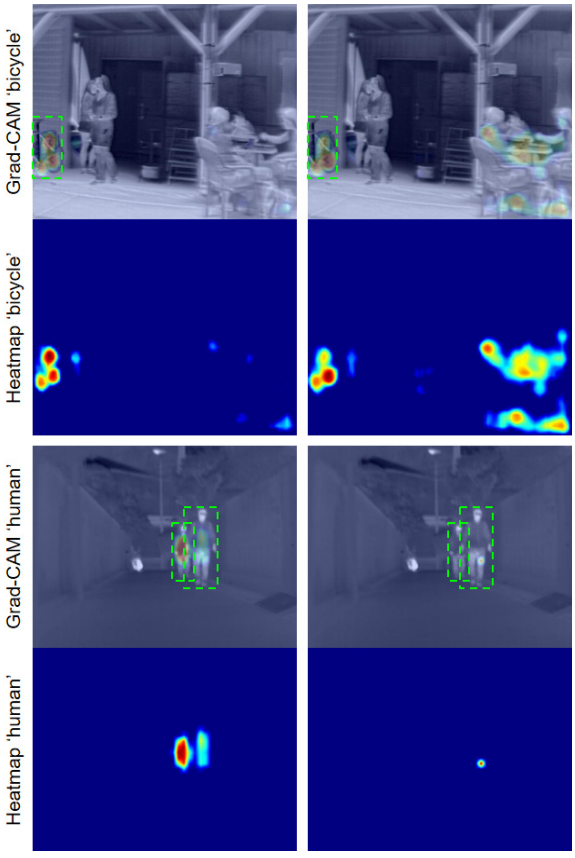


Fig. 6. AP difference $\triangle_{AP}$ comparison according to adversarial attack threshold ϵ on FLIR dataset



(a) Clean image (b) Adversarial attacks

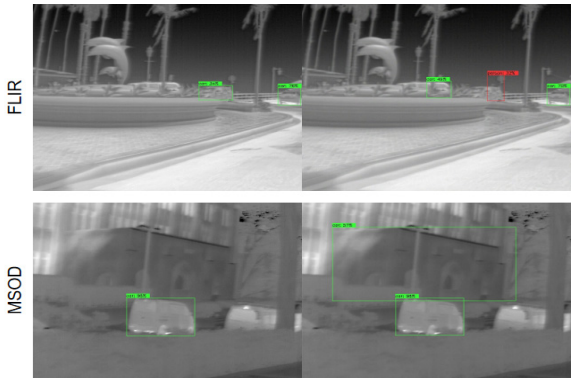
Fig. 7. Qualitative results of clean and adversarial images using Grad-CAM on FLIR and MSOD datasets

4.5 원본 영상과 적대적 공격 영상 활성화 맵 비교

제안하는 적대적 공격 방법의 효과를 정성적으로 검증하기 위해 본 실험에서는 Gradient-weighted Class Activation Mapping(Grad-CAM)[24]을 이용해 원본 영상과 적대적 영상에 대한 클래스별 활성화 맵을 비교하였다. Grad-CAM은 심층 신경망의 마지막 레이어의 경사(gradient)를 이용하여 각 클래스별로 영상의 중요한 영역을 표시한다. 이는 심층 신경망의 예측된 결과에 대해 시각적으로 설명 가능한 정보를 제공함으로써 모델의 의사 결정 이유를 이해하는 데 중요한 역할을 한다. Fig. 7에서 각 행의 첫 번째 줄의 초록색 점선 사각형은 GT에 대한 바운딩 박스를 의미하며, 각 행의 두 번째 줄은 영상의 히트맵(heatmap)을 의미한다. Fig. 7의 (a)인 원본 영상에서는 자전거와 사람을 올바르게 검출한 반면에 (b) 적대적 공격 FGSM+에서는 위쪽에서 사람과 테이블을 자전거로 오검출하였고, 아래쪽에서는 사람을 미검출하였다.

4.6 적대적 공격 예시

본 실험에서는 원본과 적대적 공격 영상에 대한 모델의 검출 결과를 비교하여 적대적 공격이 심층 신경망의 검출 기능을 효과적으로 저하시킨다는 것을 보여주고자 한다. Fig. 8의 각 행은 FLIR, MSOD 데이터셋에서의 검출 결과를 나타낸다. 먼저 첫 번째 행에서는 (a) 원본 영상의 경우 자동차(초록색)를 올바르게 검출한 것과 달리 (b) 적대적 공격 FGSM+의 경우 자동차 뒷부분을 사람(빨간색)으로 오검출하였으며, 두 번째 행의 (b)에서는 큰 건물을 자동차로 오검출하였다.



(a) Clean image (b) Adversarial attacks

Fig. 8. Detection examples of clean images and adversarial images on FLIR and MSOD datasets

5. 결론 및 토의

본 논문에서는 딥러닝 기반 적외선 객체 검출의 취약성 분석을 위해 기존의 영상 분류용 적대적 공격 기술을 확장하여 객체 검출에 활용하였으며 FLIR, MSOD와 같은 적외선 객체 검출 데이터셋에서 폭넓은 정량적, 정성적 실험을 통해 제안하는 방법의 우수성을 검증하였다. 구체적으로, 제안하는 방법은 기존 기술 대비 호환성이 높고 작은 퍼터베이션 크기만으로도 검출 정확도에 큰 악영향을 미칠 수 있는 특징을 갖는다. 본 논문을 통해 육안으로 구별하기 어려운 수준의 퍼터베이션으로 적외선 객체 검출을 위한 심층 신경망의 성능에 큰 타격을 줄 수 있다는 것을 입증하였다.

한편 적대적 공격뿐만 아니라 적대적 방어에 대해서도 활발히 연구되고 있으며, 적대적 학습 기반의 적대적 방어에 관한 연구를 후속 연구로 진행하고자 한다.

References

[1] J. Gu et al., “Recent Advances in Convolutional Neural Networks,” Pattern Recognition, Vol. 77, pp. 354-377, 2018.

[2] I. Goodfellow et al., “Explaining and Harnessing Adversarial Examples,” International Conference on Learning Representations, 2015.

[3] C. Szegedy et al., “Intriguing Properties of Neural Networks,” International Conference on Learning Representations, 2014.

[4] S. Ren et al., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, pp. 1137-1149, 2017.

[5] J. Deng et al., “ImageNet: A Large-Scale Hierarchical Image Database,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, 2009.

[6] T. Lin et al., “Microsoft COCO: Common objects in context,” European Converence on Computer Vision (ECCV), pp. 740-755, 2014.

[7] M. Tan et al., “EfficientDet: Scalable and Efficient Object Detection,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10778-10787, 2020.

[8] A. Bochkovskiy et al., “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint arXiv:2004.10934, 2020.

[9] T. Karasawa et al., “Multispectral Object Detection for Autonomous Vehicles,” Proceedings of the on Thematic Workshops of ACM Multimedia, pp. 35-43, 2017.

[10] D. Feng et al., “Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges,” IEEE Transactions on Intelligent Transportation Systems, pp. 1-20, 2020.

[11] “FREE FLIR Thermal Dataset for Algorithm Training,” FLIR Systems. Accessed April 1, 2021. https://flir.com/oem/adas/adas-dataset-form/.

[12] J. Rony et al., “Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4322-4330, 2019.

[13] A. Kurakin et al., “Adversarial Examples in the Physical World,” International Conference on Learning Representations(Workshop Track Proceedings), 2017.

[14] A. Mardy et al., “Towards Deep Learning Models Resistant to Adversarial Attacks,” International Conference on Learning Representations, 2018.

[15] U. Jang et al., “Objective Metrics and Gradient Descent Algorithms for Adversarial Examples in Machine Learning,” Proceedings of the Annual Computer Security Applications Conference, pp. 262-277, 2017.

[16] H. Hosseini et al., “On the Limitation of Convolutional Neural Networks in Recognizing Negative Images,” IEEE International Conference on Machine Learning and Applications, pp. 352-358, 2017.

[17] T. Miyato et al., “Distributional Smoothing with Virtual Adversarial Training,” International Conference on Learning Representations, 2016.

[18] S. Moosavi-Dezfooli et al., “DeepFool: A Simple

and Accurate Method to Fool Deep Neural Networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2574-2582, 2016.

[19] J. Rauber et al., "Fast Differentiable Clipping-Aware Normalization and Rescaling," arXiv preprint arXiv: 2007.07677, 2020.

[20] N. Carlini et al., "Towards Evaluating the Robustness of Neural Networks," IEEE Symposium on Security and Privacy, pp. 39-57, 2017.

[21] C. Xie et al., "Adversarial Examples for Semantic Segmentation and Object Detection," Proceedings of the IEEE International Conference on Computer Vision, pp. 1369-1378, 2017.

[22] Y. Li et al., "Robust Adversarial Perturbation on Deep Proposal-based Models," British Machine Vision Conference, 2018.

[23] X. Liu et al., "DPatch: An Adversarial Patch Attack on Object Detectors," AAAI Workshop on Artificial Intelligence Safety, 2019.

[24] R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626, 2017.