文章编号：2097-1974(2025)04-0038-07 DOI：10.7654/j.issn.2097-1974.20250405

基于补丁对抗攻击的伪装欺骗技术研究

杨 威，李晟嘉，邵子航，黄 虎，郑本昌
（中国运载火箭技术研究院研究发展中心，北京，100076）

摘要：随着人工智能技术的飞速发展，无人化系统的智能水平日益提高，其中智能侦察技术较为成熟且应用广泛，面向智能侦察的伪装欺骗技术研究迫在眉睫。针对上述问题，提出了一种基于补丁对抗攻击的伪装欺骗方法，采用卷积神经网络构建分类器作为攻击对象，通过设计全新的补丁生成方式和损失函数，完成目标样本的补丁攻击，能够有效地将攻击的目标样本映射到指定的错误目标类别上，并提供了针对性的评价方法及丰富的试验，验证了该方法的先进性与有效性。

关键词：人工智能；智能侦察；伪装欺骗；补丁攻击；对抗样本
中图分类号：V42；TP3 **文献标识码**：A

Adversarial Patch Attack Based Camouflage And Deception Method

YANG Wei, LI Shengjia, SHAO Zihang, HUANG Hu, ZHENG Benchang
(R&D Center, China Academy of Launch Vehicle Technology, Beijing, 100076)

Abstract: As artificial intelligence technology developing rapidly, the intelligence level of unmanned systems is much increasing. Specially, intelligent reconnaissance technology is more mature and widely used. To solve the above problems, an adversarial patch attack based camouflage and deception method is proposed. Convolutional neural network is used to build a classifier as the attack object, and a novel patch generation method and loss function are designed to attack target samples, which effectively maps the attacked target samples to the specified wrong target category. A directed evaluation method and wealthy experiments are provided to verify the advancement and effectiveness of this method.

Keywords: artificial intelligence; intelligent reconnaissance; camouflage and deception; patch attack; adversarial examples

0 引 言

补丁对抗攻击是人工智能安全与对抗技术中的重要攻击手段。通过改变目标样本部分区域的像素值，产生一定的对抗扰动，误导智能侦察系统，使得分类模型以较高的置信度输出错误的结果，破坏智能侦察系统的识别能力，实现对高价值目标的保护。在对智能侦察系统的伪装中，使用对抗补丁攻击，可以诱导智能侦察系统将目标错误分类到特定类别，实现对伪装对象的伪装保护。在对智能侦察系统的欺骗中，使用对抗补丁攻击，可以引导智能侦察系统将目标错误分类，提高伪装对象的生存能力。

在介绍对抗补丁攻击之前，首先概述常见的对抗样本攻击方法。在对抗样本迁移能力方面，Dong等人[1]将动量法引入到对抗样本的迭代生成过程中，提出了MI-FGSM（Momentum I-FGSM），提高了对抗样本的攻击可迁移性，对白盒模型和黑盒模型都具有更好的攻击效果。Lu等[2]充分利用模态交互，结合跨模态指导与对齐保持增强，生成对抗样本能够在多种下游视觉语言任务中的视觉模型间有效迁移。Zhu等[3]通过空域和频域特征合并到对抗生成架构中生成对抗样本，并设计了一种具有高频组件的跳跃连接的编解码器架构，以保留细粒度特征，其生成的样本能够同时欺骗分类和分割模型，证明该方法在不同任务间具有较好的迁移性。此外，针对对抗样本生成效率方面，Reza等[4]提出一种基于几何决策的黑盒攻击，通过查询高效和曲率感知，有效处理任意决策边界，特别在低曲率制作高质量对抗样本方面效果突出。同时，对抗鲁棒性方面，Huang等[5]提取领域知识占据图像的中/高频，以缩小对抗扰动的优化空间，强化预训练视觉Transform-

收稿日期：2024-11-28；修回日期：2025-04-02

er的对抗鲁棒性。

补丁对抗攻击是对抗样本攻击的一种重要手段，能够更好地进行真实世界的攻击。将对抗样本图像打印出来粘贴在目标物体上，使得神经网络无法正确分类。Brown等[6]提出Adversarial Patch，通过训练产生一个独立于拍摄光照、角度等场景的遮罩补丁，替换图像中的一部分，使分类器产生误分类。DPatch[7]将Fast-RCNN和YOLO等模型也加入到攻击目标中，能够进行有针对性和无针对性的攻击。实际物品上进行补丁对抗的研究也越来越多，一些研究[8-10]发现在衣服上打印出补丁块，穿上带有对抗补丁的衣服后能够躲避监控摄像头的追踪。Sharif等[11]设计了眼镜补丁，该设计能够破坏人脸识别系统。Song等[12-13]研究补丁对抗攻击干扰车载摄像头对交通标识的识别。Wang等[14]将补丁对抗图案喷涂在汽车表面，降低目标识别系统的识别成功率。在军事上，使用补丁对抗[15-16]隐藏高价值军用飞机和舰船，使其不受无人机上智能侦察系统的识别。目前，大部分方案仅仅考虑针对特定对象的对抗补丁攻击，泛化性不足，生成的补丁无法在不同类别目标上达到相同的效果。

本文提出了一种补丁对抗攻击的伪装欺骗方法，能够在较少计算资源下，快速在不同类别的图像上生成对抗补丁，诱导VggNet[17]和ResNet[18]等分类器将多种目标分类到指定的错误类别，起到了对智能侦察系统的伪装欺骗，实现对关键目标的保护。本文提出了一种新的对抗补丁生成方式，同时设计了新的损失函数指导具有泛化性的对抗补丁的生成。提出新的评价指标：准确率各相关指标和对抗攻击失真度指标，进行综合评价。通过对照消融试验，验证了所提的对抗补丁生成方法的有效性。

1　算法原理

1.1　分类器

卷积神经网络是深度学习特征提取的基座，在图像分类等底层智能任务中较为成熟。因此在对抗样本任务中，通常攻击对抗由卷积神经网络构建的分类器，从底层直接干扰深度特征，致使上层的人工智能技术失效。本文针对较为常用的VggNet和ResNet两种经典卷积神经网络，构建两种分类器，以生成攻击补丁并验证补丁攻击效果。其中VggNet分类器以VGG-16卷积网络部分为基础，构建新的全连接层，映射到目标分类类别，如图1所示。

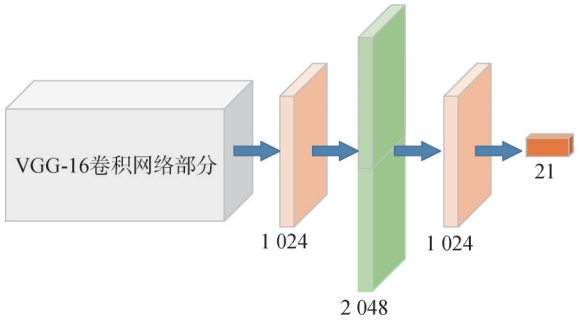

图1　VggNet分类器结构示意

Fig.1　Illustration of VggNet classifier structure

ResNet分类器以ResNet-50残差块网络部分为基础，构建新的全连接层，映射到目标分类类别，如图2所示。

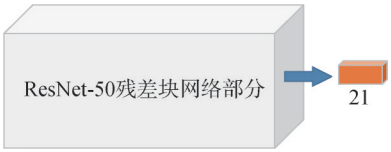

图2　ResNet分类器结构示意

Fig.2　Illustration of ResNet classifier structure

1.2　补丁生成

首先，随机生成攻击补丁进行初始化，将初始化生成的补丁与输入的图像相加，获得初始化对抗样本。再将初始化对抗样本输入到分类器中进行训练，获得梯度信息，利用梯度信息计算生成新的攻击补丁，并循环上述过程不断地更新攻击补丁。生成攻击补丁的计算分为2个主要部分：关键区域项、动量梯度项。

关键区域项是通过以当前时刻t的梯度信息中最大的变化量为基准，计算得到梯度信息中各梯度变化的程度，并根据变化程度判断图像中的关键区域，具体计算公式如下：

$$G_{\text{key}}^{t} = G_{\text{adv}}^{t} / \max\left(G_{\text{adv}}^{t}\right) \tag{1}$$

式中　G_{key}^{t}为关键区域项；G_{adv}^{t}为梯度信息。

动量梯度项则是以当前时刻t的梯度信息为基础，计算梯度方向的速度矢量，以提高对抗样本的迁移性，并且为了进一步提升梯度扰动效果，保留了梯度的符号和数值。同时引入当前时刻t的梯度信息，旨在保留原始的梯度信息，以丰富攻击补丁的生成信息，并且通过权重控制梯度信息的影响程度，具体计算公式如下：

$$G_{t}' = \varepsilon \cdot G_{\text{adv}}^{t} + G_{\text{adv}}^{t} / \left\| G_{\text{adv}}^{t} \right\| \tag{2}$$

式中 G_t'为动量梯度项；ε表示权重。

最终，将上一时刻t–1的攻击补丁、关键区域项、动量梯度项进行计算，生成当前时刻t的攻击补丁，具体计算公式如下：

$$p_t = p_{t-1} - \delta \cdot G_{\text{key}}^t + \eta \cdot G_t' \tag{3}$$

式中 p为攻击补丁；δ，η表示权重。

1.3 损失函数

本算法旨在通过攻击目标，使其无法正确分类且错误分类为指定目标，以达成伪装欺骗的效果。因此，设计了一种损失函数，通过利用该分类器输出的各类别分数向量，并将输出的分数向量中所对应的指定错误目标类别分数和正确目标类别分数作为损失函数的吸引项和排斥项，从而增大指定错误目标类别分数，并减小正确目标类别分数。具体计算公式如下：

$$L = L_{\text{target}} + \alpha \cdot L_{\text{ori}} \tag{4}$$

式中 L_{target}为吸引项；L_{ori}为排斥项；α表示权重。

1.4 补丁攻击

为了提升攻击样本的多样性和迁移性，本算法将生成的攻击补丁，通过掩码截取相应的形状（例如矩形、圆形）和尺寸，并按照随机位置、随机旋转角度、扰动范围限制等方式对攻击补丁进行处理，后将处理后的攻击补丁与原始图像相加，以完成对图像样本的攻击，最终生成目标图像的对抗样本。具体计算公式如下：

$$X_{\text{adv}} = M \odot patch + X_{\text{ori}} \tag{5}$$

式中 X_{adv}为对抗样本图像；X_{ori}为原始图像；M为掩码；$\odot$表示对应元素相乘。

1.5 算法流程

本算法包含分类器训练模块、生成攻击模块和对抗测试模块，算法流程如图3所示。

a）分类器训练模块。将正样本、负样本和对抗正样本作为输入，其中对抗正样本可由本算法生成。首先对输入的三种样本均进行预处理，再将预处理后的样本数据输入到分类器中训练，使正样本和对抗正样本映射分类到相同的标签，负样本映射分类到其本身对应的标签，并保存训练完成的分类器参数模型。

b）生成攻击模块。将正样本输入到训练完成的分类器参数模型中，针对分类正确的正样本，通过补丁生成损失函数和补丁生成算法，迭代训练生成攻击补丁，并完成对全部正样本的攻击

c）对抗测试模块。将全部生成的对抗正样本输入到由两种分类器构成的联合分类器中进行测试。若分类正确，表示该对抗正样本攻击无效；若分类不正确，表示该对抗正样本攻击有效，具备欺骗效果；若分类为指定负样本类别，表示该对抗正样本攻击有效，具备伪装效果。

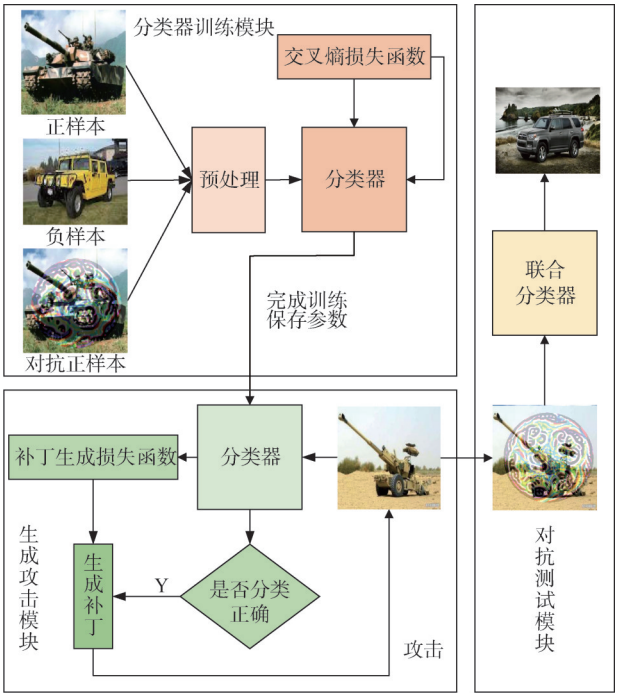

图3 算法流程

Fig.3 Flowchart of proposed algorithm

2 算法训练

2.1 数据集

本文采用国防科技大学和中国指挥与控制学会主办的“慧眼行动”全国智能算法对抗挑战赛的数据集，开展了大量试验，该数据集包含20个目标类别，分别为直升机、客机、预警机、运输机、两栖攻击舰、主战坦克、榴弹炮等，其中训练集共26 200张图像，测试集共1 200张图像。此外，针对伪装欺骗的应用目的，本文从公认的民用车辆Stanford Cars数据集[19]中分别随机选取了1 450张作为训练集和169张作为测试集，对赛事数据集进行增广。通过补丁攻击分类器，使20个目标类别被误分类为民用车辆，达到伪装欺骗的目的。此外，本文将补丁攻击之后生成的1 200张对抗样本作为训练集，以提升分类器的对抗防御能力。因此，本文采用的总数据集包含赛事数据集、Stanford Cars数据集和对抗样本数据集，分别为正样本、负样本和对抗正样本。

2.2　数据预处理

本文采用随机擦除的数据增强方式对正样本、负样本和对抗正样本图像进行数据预处理。随机擦除的数据增强方式即通过一定概率的随机擦除，使原始图像中随机面积的区域被一个布满随机值的随机尺寸矩形框覆盖，以达到训练数据集增强的目的。

2.3　参数设置

本算法在1块2080ti的GPU上对算法模型进行训练。数据预处理设置25%的概率对原始图像进行随机擦除，擦除的区域覆盖面积占比范围为［5%，25%］，擦除区域的长宽比范围为［0.3，3.3］，采用Adam优化器[20]，设置batch size为64，并使用交叉熵损失函数训练分类器。补丁生成训练过程中，设置batch size为1，输入顺序设置为shuffle，补丁形状为圆形、位置随机、360°随机旋转方向、补丁面积占比不超过0.6、补丁扰动大小范围为［−60，60］。补丁生成和损失函数中的权重设置分别为$\varepsilon = 0.5$、$\delta = 20$、$\eta = 1$、$\alpha = 0.25$。

3　试验分析

3.1　评价指标

3.1.1　准确率

准确率（Accuracy）是评价分类效果最经典的指标之一，统计分类正确的样本个数和样本总个数，通过比值计算得到准确率A。具体计算公式如下：

$$A = n/N \tag{6}$$

式中　n为分类正确的样本个数；N为样本总个数。本文涉及的准确率包括补丁攻击前分类器的准确率A_{b}；补丁攻击后分类器的准确率A_{a}，$1 - A_{\mathrm{a}}$表示欺骗的效果；未误分类指定目标的准确率A_{t}，$1 - A_{\mathrm{t}}$表示伪装效果。

3.1.2　对抗攻击失真度

对抗攻击失真度（Average Lp Distortion，ALD_p）采用L_p norm距离（p=0，2，∞）作为评价的失真度量，具体来说，L_0表示攻击后发生改变的像素数量；L_2表示原始示例和攻击示例之间的欧氏距离；L_∞表示对抗样本全维度下最大变化量。A_{LD}_p为所有攻击成功的对抗样本的平均归一化失真度，A_{LD}_p越小，对抗样本的不可感知性越强。具体计算公式如下：

$$A_{LD}\_p = \frac{1}{n}\left\| X_i^a - X_i \right\|_p \Big/ \left\| X_i \right\|_p \tag{7}$$

式中　n为样本个数；X_i^a为攻击后的样本；X_i为原始样本。本文将采用A_{LD}_2越小作为评价对抗攻击失真度的指标。

3.1.3　综合指标

综合指标（Average）是对准确率各相关指标和对抗攻击失真度指标的综合评价，包括补丁攻击后分类器的准确率A_{a}、未误分类指定目标的准确率A_{t}和对抗攻击失真度A_{LD}_2。具体计算公式如下：

$$\begin{aligned} A_{ve} &= \frac{\left[(1-\beta)\cdot(1-A_{\mathrm{a}}) + \beta\cdot(1-A_{\mathrm{t}}) + (1-A_{LD}\_2)\right]}{2} \\ &= \frac{\left[2 - A_{\mathrm{a}} + \beta\cdot(A_{\mathrm{a}} - A_{\mathrm{t}}) - A_{LD}\_2\right]}{2} \\ &= 1 - \frac{A_{\mathrm{a}} + A_{LD}\_2 - \beta\cdot(A_{\mathrm{a}} - A_{\mathrm{t}})}{2} \\ &= 1 - \frac{A_{\mathrm{a}} + A_{LD}\_2 + \beta\cdot\left|A_{\mathrm{a}} - A_{\mathrm{t}}\right|}{2} \end{aligned} \tag{8}$$

式中　β表示评价欺骗和伪装效果一致性效果的重要程度。

3.2　试验结果

首先，本文选取补丁攻击前分类器的准确率A_{b}、补丁攻击后分类器的准确率A_{a}、未误分类指定目标的准确率A_{t}、对抗攻击失真度A_{LD}_2、综合指标A_{ve}等指标进行试验结果对比分析，并设置β=0.5。其中设置攻击补丁面积占比为｛0.1，0.2，0.3，0.4，0.5，0.6｝，攻击补丁扰动大小范围为［−m，m］，m∈｛10，20，30，40，50，60｝。此外，本文开展了两种分类器（VggNet和ResNet）生成攻击补丁分别对联合分类器攻击的试验验证，并对攻击试验指标进行对比分析，效果对比结果如表1所示。

表1　两种分类器补丁攻击效果对比（β=0.5）

Tab.1　Comparison of patch attack effect by two classifiers（β=0.5）

方法	A_{b}	A_{a}	A_{t}	A_{LD}_2	A_{ve}
本文方法（VggNet）	0.957	0.060	0.575	0.282	**0.701**
本文方法（ResNet）	0.957	0.273	0.711	0.278	0.615
GDPA	0.957	0.229	0.884	0.194	0.625

注：加粗数值表示最优结果。

由表1可知，首先列举了本方法分别采用VggNet和ResNet分类器的最佳攻击结果指标，其中联合分类器的原始分类准确率为95.7%，可见VggNet分类器产生的攻击补丁具有明显的优势。具体的，VggNet的补丁攻击使联合分类器的准确率下降到了6%，具备明显的欺骗效果；同时将38.2%的目标误分类为指定目标，具备良好的伪装效果。同时，在相同参数配置下，通过与基于VggNet补丁生成攻击的先进方法GDPA[21]对比，本方法在欺骗效果及伪装效果的指标上分别有16.9%与30.9%的优势，综合指标亦高出7.6%。

此外，本文提供了不同参数设置条件下，各项指标结果的对比并进行分析，对比结果如表2所示。

表2 综合指标(A_{ve})对比(β=0.5)

Tab.2 Comparison of average (A_{ve})(β=0.5)

分类器类型	补丁面积	10	20	30	40	50	60
VggNet	0.1	0.617	0.607	0.599	0.591	0.582	0.575
	0.2	0.612	0.601	0.592	0.580	0.573	0.567
	0.3	0.609	0.592	0.589	0.584	0.588	0.574
	0.4	0.607	0.595	0.592	0.604	0.606	0.621
	0.5	0.604	0.597	0.602	0.617	0.644	0.676
	0.6	0.602	0.598	0.619	0.646	0.687	**0.701**
ResNet	0.1	0.505	0.495	0.488	0.480	0.474	0.468
	0.2	0.502	0.489	0.480	0.472	0.467	0.467
	0.3	0.497	0.484	0.478	0.483	0.493	0.511
	0.4	0.495	0.481	0.485	0.497	0.523	0.553
	0.5	0.493	0.483	0.492	0.515	0.560	0.599
	0.6	0.492	0.482	0.492	0.533	0.573	**0.615**

注：加粗数值表示最优结果。

由表2可知，当攻击补丁扰动参数较小时，随着攻击补丁面积占比提升，综合指标的较为稳定，变化不明显；当攻击补丁扰动参数为50和60时，随着攻击补丁面积占比提升，综合指标的提升较为明显。同时综合指标与攻击补丁扰动、补丁面积占比呈现正相关的趋势。值得一提的是，VggNet分类器攻击的整体效果明显高于ResNet分类器，可见以梯度信息作为补丁生成和攻击的驱动方式，对VggNet更加友好。

补丁攻击后分类器的准确率能够体现补丁攻击后的欺骗效果，对比结果如表3所示。由表3可知，VggNet分类器在较小的面积占比和扰动大小的条件下，仍具备良好的欺骗效果。当攻击补丁面积占比和扰动参数较大时，VggNet分类器产生的攻击补丁几乎能够使联合分类器失效，甚至使联合分类器的分类准确率低于10%。

表3 补丁攻击后分类器的准确率(A_a)对比

Tab.3 Comparison of accuracy of classifier after patch attack (A_a)

分类器类型	补丁面积	10	20	30	40	50	60
VggNet	0.1	0.481	0.478	0.463	0.450	0.442	0.427
	0.2	0.483	0.459	0.431	0.408	0.381	0.353
	0.3	0.481	0.444	0.401	0.354	0.293	0.277
	0.4	0.476	0.431	0.368	0.283	0.226	0.165
	0.5	0.474	0.403	0.317	0.228	0.153	0.092
	0.6	0.473	0.389	0.266	0.169	0.096	**0.060**
ResNet	0.1	0.934	0.926	0.908	0.893	0.876	0.859
	0.2	0.925	0.909	0.876	0.845	0.808	0.760
	0.3	0.927	0.899	0.838	0.749	0.664	0.588
	0.4	0.924	0.884	0.782	0.668	0.551	0.453
	0.5	0.921	0.855	0.727	0.580	0.430	0.331
	0.6	0.915	0.84	0.698	0.500	0.382	**0.273**

注：加粗数值表示最优结果。

未误分类指定目标的准确率能够体现补丁攻击后的伪装效果，对比结果如表4所示。由表4可知，在较小的面积占比和扰动参数的条件下，两种分类器的伪装效果均表现不佳。当攻击补丁面积占比和扰动参数较大时，VggNet分类器的伪装效果提升明显。可见，攻击补丁需要同时满足大面积占比和大扰动的两个条件，才能够达到良好的伪装效果。

表4 未误分类指定目标的准确率(A_t)对比

Tab.4 Comparison of no misclassify the accuracy of the specified target (A_t)

分类器类型	补丁面积	10	20	30	40	50	60
VggNet	0.1	1.000	1.000	1.000	0.999	0.999	0.998
	0.2	1.000	1.000	0.998	0.998	0.992	0.983
	0.3	0.999	0.999	0.993	0.977	0.945	0.938
	0.4	1.000	0.997	0.977	0.919	0.880	0.803
	0.5	1.000	0.994	0.952	0.879	0.760	0.641
	0.6	1.000	0.982	0.905	0.790	0.639	**0.575**
ResNet	0.1	1.000	1.000	1.000	1.000	0.998	0.997
	0.2	1.000	1.000	1.000	1.000	0.997	0.984
	0.3	1.000	1.000	1.000	0.997	0.975	0.956
	0.4	1.000	1.000	0.999	0.985	0.930	0.849
	0.5	1.000	0.999	0.997	0.965	0.863	0.749
	0.6	1.000	1.000	0.994	0.941	0.828	**0.711**

注：加粗数值表示最优结果。

对抗攻击失真度能够体现补丁攻击后的图像与原始图像之间的失真程度，对比结果如表5所示。由表5可知，随着攻击补丁面积占比和扰动大小的增加，失真度亦随之增加，呈现出正相关的趋势。不难发现，攻击补丁扰动大小对失真度的影响要高于面积占比所带来的影响。

表5 对抗攻击失真度($A_{LD\_}2$)对比

Tab.5 Comparison of average L_p distortion ($A_{LD\_}2$)

分类器类型	补丁面积	10	20	30	40	50	60
VggNet	0.1	0.024	0.047	0.070	0.093	0.116	**0.138**
	0.2	0.034	0.068	0.101	0.134	0.167	0.199
	0.3	0.042	0.084	0.125	0.166	0.207	0.245
	0.4	0.084	0.096	0.144	0.191	0.234	0.274
	0.5	0.055	0.109	0.162	0.213	0.255	0.281
	0.6	0.059	0.118	0.176	0.228	0.259	0.282
ResNet	0.1	0.023	0.047	0.070	0.093	0.115	**0.137**
	0.2	0.034	0.068	0.101	0.133	0.165	0.194
	0.3	0.042	0.084	0.124	0.161	0.194	0.222
	0.4	0.048	0.098	0.140	0.179	0.214	0.242
	0.5	0.055	0.107	0.155	0.198	0.235	0.263
	0.6	0.059	0.116	0.166	0.213	0.250	0.278

注：加粗数值表示最优结果。

3.3 消融试验

本文针对提出的攻击补丁生成方法和损失函数开展了消融试验，以证明提出方法的有效性。如表6所示，当本算法消除损失函数中的吸引项时，即不考虑将目标误分类为指定目标，从指标上不难发现，提出的方法失

去了伪装能力；当本算法消除损失函数中的排斥项时，即仅考虑将目标误分类为指定目标，从指标上可见，本方法的欺骗和伪装能力均有所下降；当本算法消除所提出的攻击补丁生成方法，并替换为较为经典的梯度动量方法时，不仅综合指标等明显下降，而且直接使本方法失去了伪装能力，可见生成攻击补丁时，保留一定的梯度信息能够提升算法的伪装效果。

表6　消融试验对比(β=0.5)

Tab.6　Comparison of ablation study(β=0.5)

方法	A_a	A_t	A_{LD_2}	A_{ve}
本文方法	**0.060**	**0.575**	**0.282**	**0.701**
本文方法-L_{target}	0.077	1.000	0.346	0.558
本文方法-L_{ori}	0.081	0.602	0.266	0.696
本文方法-P_G	0.089	1.000	0.345	0.556

注：加粗数值表示最优结果。

3.4　进一步分析

为了进一步探讨攻击补丁的面积占比及其分布对攻击效果所带来的影响，本文在最佳攻击效果的参数条件下开展试验，将面积占比0.6和扰动大小为60的圆形四等分为四个面积占比为0.15和扰动大小为60的圆形攻击补丁，生成攻击补丁并完成样本攻击。如表7所示，四等分攻击补丁的效果均有所下降，欺骗效果下降了6.4%，伪装效果下降了18.2%，综合指标下降了6.6%。具体攻击补丁样本对比如图4所示。

表7　各项指标对比(β=0.5)

Tab.7　Comparison of various indicators(β=0.5)

分类器	A_b	A_a	A_t	A_{LD_2}	A_{ve}
VggNet	0.957	**0.060**	**0.575**	**0.282**	**0.701**
VggNet-split	0.957	0.124	0.757	0.290	0.635

注：加粗数值表示最优结果。

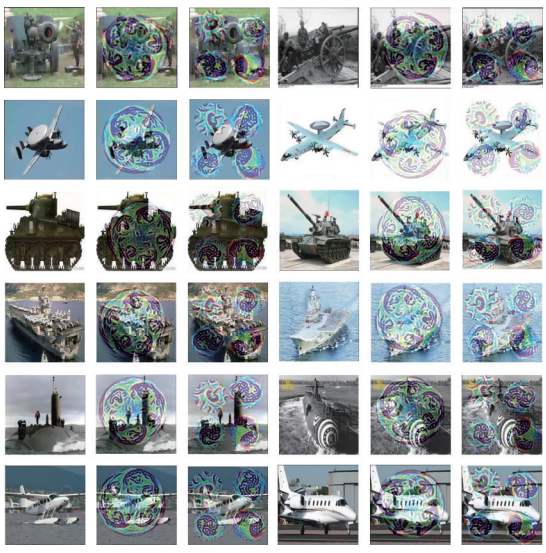

图4　两种攻击补丁效果示意

Fig.4　Illustration of both attack patch effect

此外，进一步探索本算法的拓展应用，将补丁攻击分类任务拓展到目标检测与识别任务，本文选取公认的目标检测与识别数据集2016年高分辨率船舶数据集（HRSC2016）[22]、主流算法Yolov5[23]及评价指标，以验证本算法攻击效果的普适性。此外，考虑目标检测与识别任务存在目标数量多、分布广、像素少等特点，本算法生成的攻击补丁所占比重小于目标检测框面积的2%，设置信度阈值为0.7。如表8所示，补丁攻击后Yolov5的精确率（P）、召回率（R）和平均精度值（mAP@0.5、mAP@0.5：0.95）分别下降了25.3%、17.5%、22.9%、18.2%。如图5所示，本算法能够有效降低目标检测的识别置信度，同时提升检测错误目标的概率，从而使目标检测与识别算法无法正确检测识别目标。

表8　目标检测与识别任务补丁攻击效果对比

Tab.8　Comparison of object detection and recognition

方法	Class	P	R	mAP@0.5	mAP@0.5:0.95
Yolov5	boat	0.522	0.337	0.397	0.297
Yolov5 w/本文	boat	**0.269**	**0.162**	**0.168**	**0.115**

注：加粗数值表示最优结果。

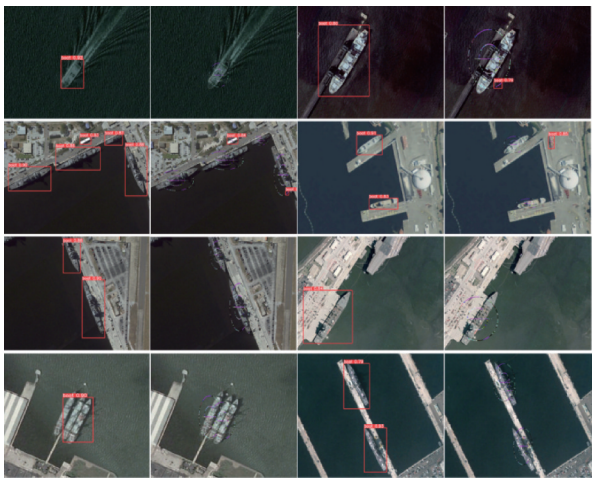

图5　攻击目标检测与识别算法效果示意

Fig.5　Illustration of attack object detection and recognition effect

4　结　论

为了应对智能侦察技术的日益成熟及广泛应用，本文提出了一种基于补丁对抗攻击的伪装欺骗方法，设计了全新的补丁生成方式和损失函数，以及对应的评价方法。面向欺骗和伪装两个维度，提供了丰富的试验对比分析，能够通过攻击目标样本，使目标具备误分类的欺骗能力，甚至误分类为指定目标的伪装能力，证明了所提方法的合理性和有效性。

参 考 文 献

[1] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]. Salt Lake City: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[2] LU D, WANG Z, WANG T, et al. Set-level guidance attack: boosting adversarial transferability of vision-language pre-training models[C]. Paris: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[3] ZHU P, OSADA G, KATAOKA H, et al. Frequency-aware GAN for adversarial manipulation generation[C]. Paris: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[4] REZA M F, RAHMATI A, WU T, et al. Cgba: curvature-aware geometric black-box attack[C]. Paris: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[5] HUANG Q, DONG X, CHEN D, et al. Improving adversarial robustness of masked autoencoders via test-time frequency-domain prompting[C]. Paris: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[6] BROWN T B, DANDELION Mané, ROY A, et al. Adversarial Patch[EB/OL]. (2017-11-27) [2024-11-10]. https://arxiv. org/abs/1712.09665.

[7] LIU Xin, YANG Huanrui, LIU Ziwei, et al. Dpatch: an adversarial patch attack on object detectors[EB/OL]. (2019-04-23)[2024-11-10]. https://arxiv.org/abs/1806.02299.

[8] THYS S, RANST W V, GOEDEME T. Fooling automated surveillance cameras: adversarial patches to attack person detection[J]. IEEE, 2019. DOI: 10.1109/CVPRW.2019.00012.

[9] HUANG L F, GAO C Y, ZHOU Y Y, et al. Universal physical camouflage attacks on object detectors[C/OL]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[10] WU Z X, LIM S N, DAVIS L S, et al. Making an invisibility cloak: real world adversarial attacks on object detectors[C]//Computer Vision-ECCV 2020. Cham: Springer, 2020: 1-17.

[11] SHARIF M, BHAGAVATUL S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of the-art face recognition[C]. New York: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.

[12] SONG D, EYKHOLT K, EVTIMOV I, et al. Physical adversarial examples for object detectors[C]. Berkeley: 12th USENIX Workshop on Offensive Technologies, 2018.

[13] CHEN S T, CORNELIUS C, MARTIN J, et al. Shape shifter: robust physical adversarial attack on faster R CNN object detector[C]// European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2019: 52-68.

[14] WANG J K, LIU A S, YIN Z X, et al. Dual attention sup pression attack: generate adversarial camouflage in physical world[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 8561-8570.

[15] AURDAL L, LKKEN K H, KLAUSEN R A, et al. Adversarial camouflage for naval vessels[C]. Bellingham: Artificial Intelligence and Machine Learning in Defense Applications, 2019.

[16] ADHIKARI A, HOLLANDER R D, TOLIOS I, et al. Adversarial patch camouflage against aerial detection[C]. Bellingham: Artificial Intelligence and Machine Learning in Defense Applications II, 2020.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. San Diego: International Conference on Learning Representations, 2015.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. IEEE, 2016. DOI: 10.1109/CVPR.2016.90.

[19] KRAUSE J, STARK M, DENG J, et al. 3D object representations for fine-grained categorization[C]. Sydney: IEEE International Conference on Computer Vision Workshops, 2014.

[20] KINGMA D, BA J. Adam: A method for stochastic optimization[EB/OL]. (2017-11-29)[2024-11-10]. https://arxiv.org/abs/1412.6980.

[21] LI X, JI S B. Generative dynamic patch attack[EB/OL]. (2021-11-15) [2024-11-10]. https://arxiv.org/abs/2111.04266.

[22] LIU Z, YUAN L, WENG L, et al. A high resolution optical satellite imagedataset for ship recognition and some new baselines[C]. Hangzhou: Chinese Conference on Pattern Recognition and Machine Learning, 2017.

[23] ZHAN W, SUN C, WANG M, et al. An improved Yolov5 real-time detection method for small objects captured by UAV[J]. Soft Comput, 2022(26): 361-373.

作 者 简 介

杨 威（1981—），男，研究员，主要研究方向为智能与仿真系统、对抗博弈技术。

李晟嘉（1995—），男，工程师，主要研究方向为深度学习、图像感知、人工智能安全对抗技术。

邵子航（1999—），男，工程师，主要研究方向为强化学习、人工智能安全对抗技术。

黄 虎（1986—），男，研究员，主要研究方向为智能博弈对抗、仿真建模、人工智能安全对抗技术。

郑本昌（1986—），男，高级工程师，主要研究方向为强化学习、多智能体博弈、人工智能安全对抗技术。