Труды МАИ. 2025. № 143
Trudy MAI. 2025. No. 143. (In Russ.)

Научная статья
УДК 621.396.96
URL: https://trudymai.ru/published.php?ID=185647
EDN: https://www.elibrary.ru/BBLPAJ

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ СОСТЯЗАТЕЛЬНОЙ ONE-PIXEL-АТАКИ НА НЕЙРОННЫЕ СЕТИВ ЗАДАЧЕ СРЫВА КЛАССИФИКАЦИИ РАДИОЛОКАЦИОННЫХ ИЗОБРАЖЕНИЙ

Иван Федорович Купряшкин

Военный учебно-научный центр Военно-воздушных сил «Военно-воздушная академия имени профессора Н.Е. Жуковского и Ю.А. Гагарина» (г. Воронеж),

Воронеж, Россия

ifk78@mail.ru

Аннотация. Статья посвящена оценке возможности использования уязвимости нейросетевой системы обработки радиолокационных изображений к состязательным атакам для повышения эффективности средств активного противодействия радиолокационным системам обзора земной поверхности. В качестве нейросетевой системы обработки рассмотрены сверточные сети и сети-трансформеры при различном сочетании гиперпараметров. В качестве воздействия рассмотрена ретрансляционная помеха, обеспечивающая формирование ложной точечной отметки на радиолокационном изображении. Установлено, что возможна реализация эффективной One-Pixel-атаки, обеспечивающей энергетический выигрыш на один-два порядка по

сравнению с традиционным воздействием, однако условием этого является наличие точных сведений об архитектуре нейронной сети, используемой для обработки изображений, и точных сведений о характеристиках радиолокационной станции и местоположении ее носителя в момент съемки.

Ключевые слова: нейронная сеть, состязательная атака, радиолокационное изображение

Для цитирования: Купряшкин И.Ф. Исследование эффективности состязательной One-Pixel-атаки на нейронные сети в задаче срыва классификации радиолокационных изображений // Труды МАИ. 2025. № 143. URL: https://trudymai.ru/published.php?ID=185647

Original article

STUDY OF THE ONE-PIXEL ADVERSARIAL ATTACK ON NEURAL NETWORKS EFFECTIVENESS IN THE TASK OF DISRUPTING THE CLASSIFICATION OF RADAR IMAGES

Ivan. F. Kupryashkin

Military Educational and Scientific Center of the Air Force "N.E. Zhukovsky and Y.A. Gagarin Air Force Academy" (Voronezh),

Voronezh, Russia

ifk78@mail.ru

Abstract. Modern space radar systems are a very informative source of information, and therefore they are of considerable interest to specialists in electronic warfare as an object of

active electronic countermeasures. Given that neural networks that are sensitive to adversarial attacks are increasingly used to process radar images, it is likely that approaches to implementing countermeasures using methods based on this new vulnerability will emerge.

The paper is devoted to assessing the possibility of using the vulnerability of neural network radar images processing system to adversarial attacks to improve the effectiveness of active countermeasures to space radars. As a neural network processing system, convolutional networks and transformer networks with different combinations of hyperparameters are considered. The impact considered is a retransmitted signal that ensures the formation of a false point target on a radar image. It has been established that it is possible to implement an effective One-Pixel attack, providing an energy gain of one to two orders.

It is shown that shifting the false point target by just a few resolution elements leads to a significant decrease in the effectiveness of the attack. In addition, it was found that the One-Pixel attack is characterized by low portability, since not only a significant change in architecture (from a convolutional network to a transformer network), but also a not very significant change in hyperparameters led to an almost complete leveling of the effect of the impact.

That is, the condition for an effective adversarial One-Pixel attack is the presence of precise information about the architecture of the neural network used for image processing, and precise information about the characteristics of the radar and the location of its carrier at the time of shooting.

It is possible that some types of adversarial attacks may be less sensitive to changes in the architecture of the attacked network or to spatial displacement of the perturbation. In this regard, the issues of generating interference capable of implementing the effect of a adversarial attack on space radars require further study.

Keywords: neural network, adversarial attack, radar image

For citation: Kupryashkin I.F. Study of the One-Pixel adversarial attack on neural networks effectiveness in the task of disrupting the classification of radar images. *Trudy MAI*. 2025. No. 143. (In Russ.). URL: https://trudymai.ru/eng/published.php?ID=185647

Введение

Несмотря на впечатляющие успехи, достигнутые нейронными сетями в задачах компьютерного зрения [1-4], они характеризуются таким существенным недостатком, как уязвимость к состязательным атакам (adversarial attacks) [5-10]. Состязательная атака состоит в незначительном, практически незаметном, искажении входных данных нейронной сети, тем менее приводящим к полностью неверному результату их обработки [7, 11]. Детальная систематизация как состязательных атак, так и методов защиты от них приведена в [7-9, 12, 13].

Так как нейросети демонстрируют высокие результаты в том числе и в задачах обработки радиолокационных изображений (РЛИ) [14-18], естественно, что вопросы оценки их уязвимости к состязательным атакам также привлекли внимание

исследователей [12, 13, 19-21]. В целом отмечается достаточно высокая чувствительность нейросетей даже к незначительному изменению РЛИ.

Следует отметить, что радиолокационные системы космического базирования вообще являются одним из важнейших источников данных о земной поверхности благодаря высокой разрешающей способности, обеспечению которой уделяется значительное внимание [22, 23].

И это, в свою очередь, обусловливает важность космических РЛС как объектов активного противодействия, направленного на скрытие заданных районов земной поверхности от детальной радиолокационной съемки [21]. Активное противодействие РЛС состоит в создании преднамеренных помех различного типа (маскирующих шумовых [24], имитационных [25, 26] и др.), требования к энергетике которых определяются большим количеством факторов – взаимным расположением средства помех и РЛС, степенью помехозащищенности РЛС, видом формируемой помехи и целевым эффектом ее воздействия и т.п. Но, так или иначе, энергетические требования к средствам помех радиолокации обычно являются достаточно жесткими [24]. Их возможное смягчение за счет использования новых потенциальных уязвимостей системы обработки РЛС является вопросом, безусловно требующим пристального внимания.

Однако в работах [12, 13, 19-21], посвященных изучению состязательных атак на нейросетевые системы обработки РЛИ, вопросы формирования помех, результатом обработки которых в РЛС является эффективное состязательное возмущение, не

рассматриваются. Приводятся синтезированные различными способами примеры этих возмущений, однако совершенно не комментируется возможность их реального формирования в системе обработки РЛС за счет специального помехового воздействия стороннего источника, то есть средства помех.

В связи с этим *целью работы* является оценка возможности формирования и эффективности преднамеренных помех, эффект воздействия которых основан на уязвимости нейросетевых алгоритмов обработки радиолокационных изображений к состязательным атакам.

Состязательные атаки обычно объединяют в две группы: White Box (WB) и Black Box (BB) [7-9]. В случае WB в распоряжении атакующего имеется полная информация о сети, включая значения всех ее весов. В случае BB доступ имеется только ко входу и выходу сети, так же могут иметься полные или частичные сведения об используемой в ней архитектуре. Атаки также классифицируются на Targeted и Non Targeted [7-9]. Targeted-атаки подразумевают достижение конкретного желаемого результата воздействия. Например, целью такой атаки может являться конкретный класс (отличный от истинного), к которому нейронная сеть должна отнести подаваемый на ее вход пример. В случае Non Targeted-атаки целью является собственно неправильная классификация, то есть в этом случае неважно, к какому классу сеть отнесет входной пример, лишь бы не к истинному.

Специфика процесса создания помех [24, 25] РЛС обзора земной поверхности и требование наличия полного доступа к атакуемой сети пока не позволяет рассматривать

WB-атаку как реализуемую на практике, в связи с чем представляется целесообразным изучение возможности осуществления эффективной BB-атаки как требующей существенно меньшего объема информации о подавляемой РЛС и условиях противодействия. Одним из видов BB-атак на нейросети является атака One-Pixel [27], или OP-атака. Такое наименование атаки отражает ее характер в самом буквальном смысле – во многих случаях достаточно изменить интенсивность отдельного пикселя, чтобы изменить результат классификации всего изображения на неверный.

Если рассматривать процесс активного противодействия РЛС обзора земной поверхности, то в качестве эквивалента OP-атаки могут выступать ложные точечные отметки (ЛТО) на РЛИ, формируемые путем создания когерентной ретрансляционной помехи так, как это описано в [24, 25].

Создание ретрансляционной помехи РЛС обзора поверхности состоит в приеме средством помех зондирующего импульса РЛС, его запоминании и последующем воспроизведении. Если изменение времени задержки и начальной фазы переизлучаемых импульсов соответствуют закону изменения расстояния между РЛС и средством помех, то результатом их обработки в течение интервала синтезирования будет являться отметка ложного точечного объекта. Смещение отметки по дальности и азимуту относительно позиции средства помех определяется временной задержкой и вносимым доплеровским сдвигом соответственно. При воспроизведении в каждом периоде зондирования не одной, а множества копий импульса, результатом обработки в РЛС будет являться совокупность соответствующего количества ЛТО.

То есть, особенности функционирования РЛС обзора поверхности и средств помех ретрансляционного типа допускают техническую возможность ОР-атаки на РЛИ. Ниже рассматривается Non Targeted ОР-атака, так как на практике важно заставить противника принять неправильное решение не в пользу конкретного класса объекта, а в принципе ошибиться.

Основная часть

В качестве исходных данных использован набор MSTAR [18,28], из всех имеющихся изображений которого сформированы обучающий, проверочный и тестовой наборы, сведения об объемах (количестве изображений) которых приведены в таблице 1. Общее количество изображений обучающего, проверочного и тестового наборов составляет 1923, 891 и 2503 соответственно.

Таблица 1 – Характеристики обучающих, проверочных и тестовых наборов

Набор	Объект	Объем	Объект	Объем	Объект	Объем	Объект	Объем	Объект	Объем
Обучающий	2С1	209	D7	200	БРДМ-2	209	БТР-70	163	ЗиЛ-131	200
Проверочный		90		99		89		70		99
Тестовый		274		274		274		196		274
Обучающий	БМП-2	163	Т-62	200	БТР-60	179	Т-72	200	ЗСУ-23	200
Проверочный		70		99		77		79		99
Тестовый		195		273		195		274		274

Задачей алгоритма ОР-атаки является определение уровня и координат (на РЛИ) пикселя, который необходимо изменить для срыва классификации. По сути, это задача оптимизации, при которой минимизируется доверие к истинному классу. Особенность

ВВ-атаки состоит в отсутствии возможности построения гладкой целевой функции, и, соответственно, использования градиентных методов оптимизации. В связи с этим в качестве оптимизационного обычно применяется метод дифференциальной эволюции, относящийся к классу генетических алгоритмов [30]. Оптимизируемой функцией являются результаты предсказания сети по каждому классу.

В качестве исходного варианта использовалась сеть VGG-типа [2], включающая две пары сверточных слоев с ядрами 3×3, единичным шагом свертки и количеством фильтров 32 и 64 в слоях каждой пары соответственно. После каждой пары включены слои подвыборок 2×2. Входной полносвязный слой классификатора имеет 4096 входов и 256 выходов, слой 50-процентного прореживания, выходной полносвязный слой имеет десять выходов по числу классов объектов. Функция активации сверточных слоев и входного слоя классификатора – ReLU. Размерность входа 44×44, количество весов сети 1’116’394. В качестве приема предотвращения переобучения применяется расширение данных путем смещения каждого РЛИ обучающего набора по вертикали и горизонтали на случайное число пикселей от одного до пяти.

Точность на тестовом наборе (объемом 2503) составила 96,52%, то есть общее количество правильно классифицированных изображений составляет 2416.

Далее каждое из них подвергалось ОР-атакам, результативными из которых явились 978. Под результативными понимаются атаки, приведшие к неправильной классификации объекта, и, таким образом, точность работы сети снизилась до 59,52%. Координаты и уровни всех «атакующих» пикселей и результатов атак фиксировалось

для дальнейшего анализа. На рисунке 1 приведены примеры атакованных РЛИ объектов всех классов. Отчетливо наблюдаются яркие пиксели, которые могут находиться как в пределах самих отметок объектов (БМП-2, БРДМ-2, ЗиЛ-131), так и рядом с ними.

Для оценки влияния уровня пикселя на результативность атаки сформированы десять тестовых наборов по 2416 изображений в каждом. На каждом изображении уровню пикселя с координатами, определенными в ходе ОР-атаки, присваивалось значение, составляющее от 10% до 100% от исходного. В пределах каждого отдельного тестового набора относительные уровни пикселей являются одинаковыми. Примеры изображений гаубицы 2С1 из каждого набора приведены на рисунке 2, а соответствующие значения точности классификации – в таблице 2. Из полученных результатов следует ожидаемый вывод о снижении эффективности атаки по мере уменьшения ее уровня.

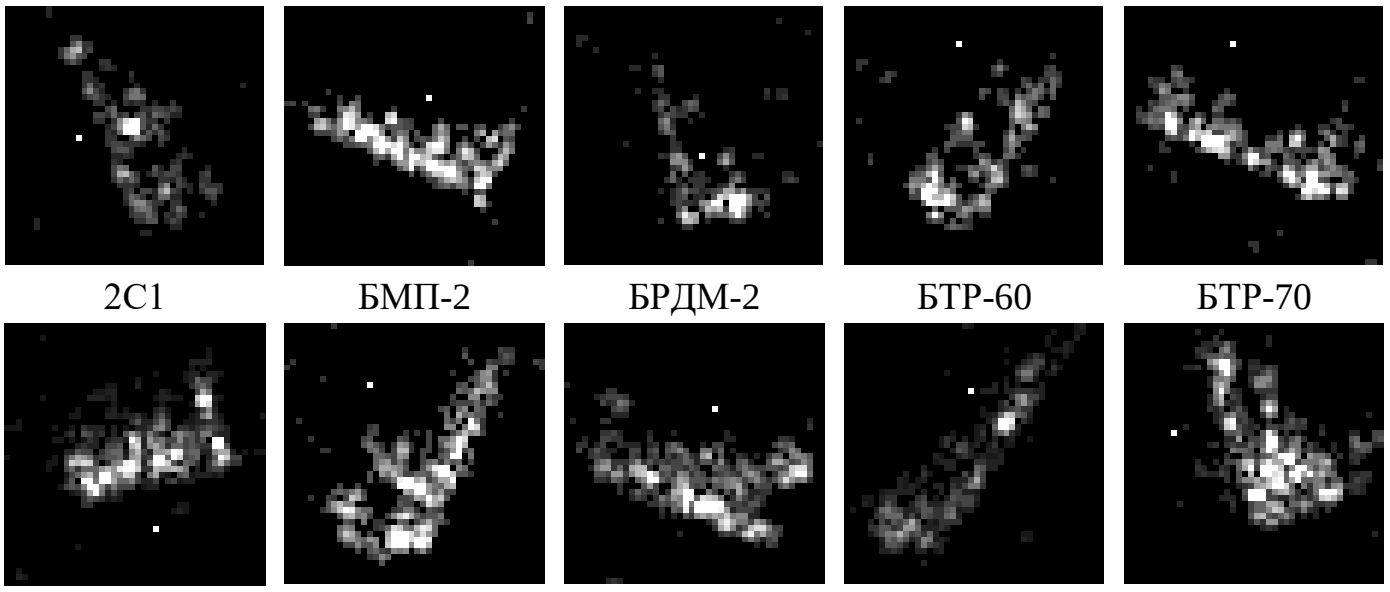


Рисунок 1 – РЛИ объектов набора MSTAR в результате применения ОР-атаки

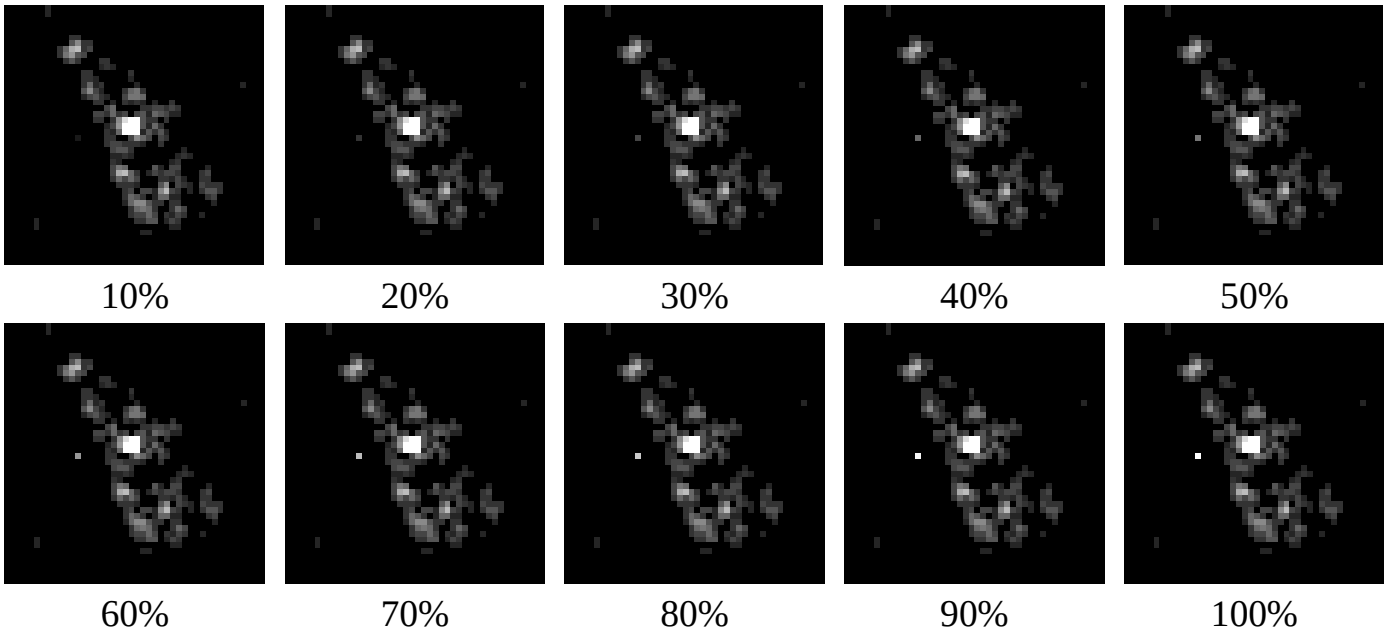


Рисунок 2 – РЛИ гаубицы 2С1 при различном уровне ОР-атаки

Таблица 2 – Точность классификации изображений набора MSTAR при различном уровне ОР-атаки

Уровень, %	10	20	30	40	50	60	70	80	90	100
Точность, %	99,21	98,14	96,65	94,12	91,18	87,67	83,94	79,39	73,09	59,52

Несмотря на то, что рассматриваемый вид атаки носит название One-Pixel, количество изменяемых пикселей изображения, вообще говоря, может быть произвольным. Результаты оценки точности классификации при различном количестве пикселей приведены в таблице 3. Кроме того, для сравнения приведены соответствующие значения точности при случайном равномерном размещении такого же количества пикселей. Из полученных результатов следует однозначный вывод о заметно большей эффективности распределения пикселей с использованием эволюционного алгоритма по сравнению с полностью случайным.

Таблица 3 – Точность классификации при различном количестве пикселей ОР-атаки

Количество пикселей	ОР-атака		Случайное размещение	
	Количество успешных атак	Точность, %	Количество успешных атак	Точность, %
1	978	59,52	51	97,89
3	1633	32,41	181	92,51
5	1795	25,70	318	86,84
9	1963	18,75	534	77,90
17	2140	11,42	933	61,38
25	2216	8,28	1191	50,70
50	2298	4,88	1655	31,50

Следует отметить, что полученные результаты соответствуют ситуации, когда изменяются отдельные пиксели изображения. Однако при формировании ретрансляционной помехи результатом ее обработки является ЛТО, сечение которой по координатам дальности и азимута определяется видом функции неопределенности сигнала [24]. Иначе, изменяется не отдельный пиксель изображения, а несколько, расположенных симметрично по дальности и азимуту относительно максимума.

Для оценки влияния этого фактора на эффективность ОР-атаки сформирован набор тестовых изображений (рисунок 3), представляющих собой сумму исходного изображения и отсчетов ЛТО вида

$$\rho_{ij} = \exp\left\{-\frac{1}{2}\left((i-X)^2 + (j-Y)^2\right)\right\}, \quad (1)$$

где $i = \overline{1, N}$; $j = \overline{1, N}$; $N = 44$; X и Y – номера строки и столбца пикселя, определенные в ходе ОР-атаки на соответствующее тестовое изображение.

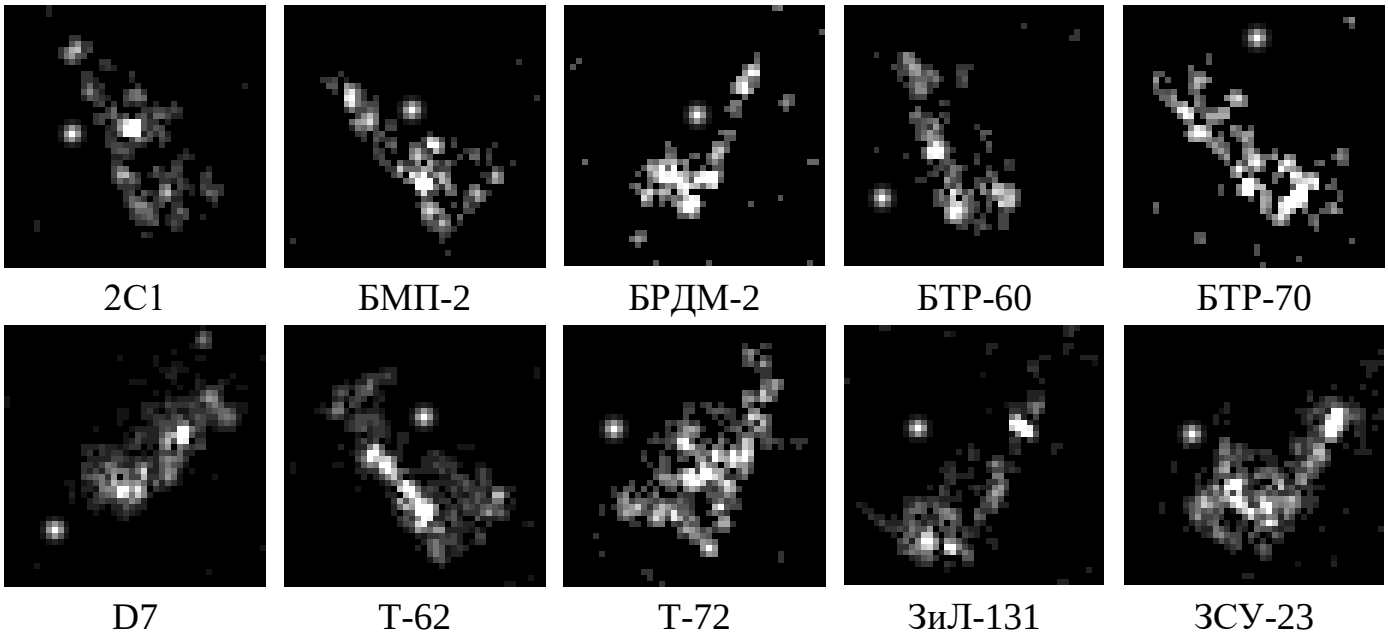
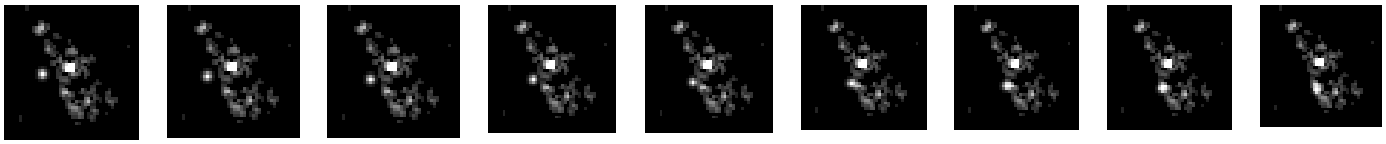


Рисунок 3 – РЛИ объектов набора MSTAR с ЛТО с максимумом, уровень и смещение которого определены по результатам ОР-атаки

Точность сверточной сети на этом наборе составила 40,56%, то есть эффективность ОР-атаки с учетом особенностей обработки помехи в РЛС даже несколько увеличилась (без учета этих особенностей точность в результате ОР-атаки снизилась только до 59,52%). В [31] показано, что аналогичный эффект при случайном расположении ЛТО достигается лишь при условии, когда их количество составляет более ста, то есть энергетический выигрыш от ОР-атаки в рассматриваемом примере составляет два порядка и более.

Формирование ЛТО в заданной точке РЛИ требует высокой точности оценки местоположения носителя РЛС относительно средства помех, а также оценок параметров сигнала [24]. Ошибки этих оценок приводят в том числе и к смещению ЛТО относительно заданного положения [24]. Для оценки влияния смещения ЛТО на

эффективность ОР-атаки дополнительно сформированы восемь наборов тестовых изображений, смещение ЛТО в которых одновременно по строке и столбцу составляет от одного до восьми пикселей. Примеры изображений гаубицы 2С1 из каждого набора приведены на рисунке 4. Результаты классификации приведены в таблице 4.



0 1 2 3 4 5 6 7 8

Рисунок 4 – РЛИ гаубицы 2С1 при различном смещении ЛТО

Таблица 4 – Точность классификации при различном смещении ЛТО

Смещение	0	1	2	3	4	5	6	7	8
Точность, %	40,56	54,01	70,86	82,41	88,45	91,85	93,42	94,25	95,41

Из полученных результатов следует, что смещение ЛТО приводит к довольно быстрому снижению эффективности ОР-атаки. Так, при смещении уже на три и более пикселя точность классификации увеличивается до 80% и более, что вряд ли можно считать эффективным результатом с точки зрения задач воздействия.

Естественно ожидать, что в реальных условиях доступ к нейронной сети, используемой противником для обработки РЛИ, будет отсутствовать. В связи с этим необходимым представляется исследование переносимости ОР-атаки, то есть ее способности к сохранению эффективности при воздействии на другие сети. Для этой цели были обучены пять дополнительных вариантов сверточной сети, сведения об архитектурах которых приведены в таблице 5. Общим для всех вариантов является применение описанной выше архитектуры с последовательным включением пар

сверточных слоев и слоев подвыборки после каждой пары. Размерность входного слоя во всех случаях составляет 44×44.

Таблица 5 – Характеристики архитектур дополнительных сверточных сетей

№ сети	Количество пар сверточных слоев	Количество фильтров в слоях каждой пары	Размерность апертуры фильтра сверточного слоя	Размерность входного/ выходного слоя классификатора	Количество настраиваемых параметров
Исход-ная	2	32/64	3	4096/256	1’116’394
1	2	16/16	3	1024/64	73’370
2	2	48/48	3	3072/256	852’090
3	2	32/64	5	1600/256	592’618
4	2	64/128	3	8192/512	4’458’954
5	4	64/128/256/512	3	18432/512	14’127’050

Точности классификации, достигнутые этими сетями на исходном тестовом наборе (до OP-атаки), на наборе, полученном по результатам OP-атаки на исходную сеть, и на наборе, сформированном в соответствии с (1), приведены в таблице 6. Из этих результатов следует вывод о достаточно низкой переносимости OP-атаки с одной сети на другую, даже при несущественных отличиях их архитектур.

Таблица 6 – Точности классификации, достигнутые сверточными сетями с отличающейся архитектурой

№ сети	Точность, %		
	Набор до OP-атаки	Набор после OP-атаки на исходную сеть	Набор после OP-атаки с учетом особенностей формирования ЛТО
Исходная	96,52	59,52	40,56
1	94,54	85,14	60,39
2	96,81	77,11	63,87
3	96,23	89,86	71,15
4	98,26	78,68	68,13
5	96,73	85,72	80,63

Одной из наиболее современных и перспективных архитектур нейронных сетей, в том числе и для решения задач обработки РЛИ на сегодняшний день считаются трансформеры [33-35]. Одной из самых первых и наиболее популярных является архитектура Vision Transformer (ViT) [35]. Разработаны два варианта трансформера с одинаковой архитектурой, но отличающимися гиперпараметрами – ViT-2D-1 и ViT-2D-2. Общими для обоих вариантов являются размерность входа 44×44, количество 16 и размерность 11×11 фрагментов разбиения входного изображения. Также общим является использование функции активации GeLU, четырех «голов» в слоях многоголового внимания, двухслойных полносвязных классификаторов с выходом на десять классов и функцией активации выхода softmax. Отличия ViT-2D-1 и ViT-2D-2 состоят: в размерности позиционного кодировщика – 96 и 128; в количестве блоков трансформера – 6 и 8; в размерности векторов ключа, запроса и значения – 72 и 96; в размерности входов полносвязных классификаторов – 1536 и 2048 соответственно.

Гиперпараметры трансформера ViT-2D-1 выбирались таким образом, чтобы общее количество весов, равное 1 136 970, примерно соответствовало количеству весов рассмотренной выше сверточной нейросети (1 116 394) для максимально объективного сопоставления результатов. Для более «тяжелой» сети ViT-2D-2 количество весов составило 2 264 330. Обучение трансформеров осуществлялось в течение шестисот эпох, для предотвращения переобучения применялся тот же прием, что и при обучении сверточной сети (случайные смещения РЛИ).

Из 2503 изображений тестового набора трансформерами ViT-2D-1 и ViT-2D-2 правильно классфицированы 2405 и 2362 изображения, то есть их точность составила 96,08% и 94,37% соответственно. В результате OP-атаки точности на тестовом наборе составили 69,77% и 72,1%, то есть архитектура трансформера продемонстрировала заметно большую устойчивость по сравнению со сверточной сетью, по крайней мере в рассматриваемом примере. Распределение координат пикселей и их уровней практически не отличается от распределения в случае сверточной сети.

Аналогично тому, как это проводилось для сверточной сети, оба варианта трансформера проверялись на устойчивость к увеличению количества пикселей при OP-атаке, а также на устойчивость к смещению пикселя по строкам и столбцам изображения.

Результаты приведены в таблицах 7 и 8. Для удобства сравнения в них также приведены уже описанные результаты, полученные сверточной сетью.

Таблица 7 – Точность классификации трансформеров при различном количестве пикселей OP-атаки

Количество пикселей	Сверточная сеть		ViT-2D-1		ViT-2D-2	
	Количество успешных атак	Точность, %	Количество успешных атак	Точность, %	Количество успешных атак	Точность, %
1	978	59,52	727	69,77	659	72,10
3	1633	32,41	1363	43,33	1419	39,92
5	1795	25,70	1569	34,76	1642	30,48
9	1963	18,75	1806	24,91	1866	21,00
17	2140	11,42	1988	17,34	2034	13,89
25	2216	8,28	2111	12,22	2125	10,03
50	2298	4,88	2214	7,94	2166	8,30

Таблица 8 – Точность классификации трансформеров при различном смещении ЛТО

Смещение		0	1	2	3	4	5	6	7	8
Точность, %	CNN	40,56	54,01	70,86	82,41	88,45	91,85	93,42	94,25	95,41
	ViT-2D-1	65,22	62,71	68,58	74,96	80,48	84,65	88,78	90,73	92,94
	ViT-2D-2	60,54	60,71	67,30	71,81	78,02	84,09	87,33	90,65	92,39

Трансформер ViT-2D-1 продемонстрировал более высокую, хотя и незначительно, устойчивость к увеличению количества пикселей ОР-атаки по сравнению с ViT-2D-2. Оба трансформера в этой ситуации также продемонстировали уверенное преимущество по сравнению со сверточной сетью. ViT-2D-1 продемонстрировал более высокую по сравнению с ViT-2D-2 устойчивость и к смещению пикселя. Тем не менее в этом отношении оба трансформера несколько проигрывают сверточной сети, так как в случае последней эффективность ОР-атаки снижается быстрее по мере увеличения смещения.

Для оценки переносимости ОР-атаки между сетями с существенно отличающимися архитектурами наборы изображений, полученные в результате ОР-атак трансформеров, подавались на вход сверточной сети, и наоборот – наборы, полученные в результате ОР-атак сверточной сети, подавались на трансформеры. Результаты, приведенные в таблице 9, свидетельствуют о практически полном отсутствии переносимости атаки между трансформерами и сверточной сетью, так как точность на «не своих» наборах во всех случаях превысила 90%.

Таблица 9 – Оценка переносимости ОР-атаки на сети с существенно различающимися архитектурами

Сеть	Атака		
	OP_{CNN}	$OP_{ViT\text{-}2D\text{-}1}$	$OP_{ViT\text{-}2D\text{-}2}$
CNN	59,52	91,5	92,13

ViT-2D-1	92,05	65,22	-
ViT-2D-2	91,02	-	60,54

Выводы

1. На примере противодействия распознаванию РЛИ объектов военной техники показано, что один из вариантов BlackBox-атаки – так называемую One-Pixel атаку – возможно осуществить за счет формирования когерентной ретрансляционной помехи, результатом обработки которой в РЛС является формирование ложной точечной отметки. Энергетический выигрыш в таком случае составляет как минимум один-два порядка по сравнению с ретрансляционной помехой, маскирующей участок местности с расположенным на нем объектом совокупностью из десятков или сотен ложных точечных отметок.

2. Однако говоря о реализуемости OP-атаки на РЛС обзора поверхности, не следует забывать о том, что и в этом случае достижение ее эффективности требует выполнения ряда условий. В первую очередь, необходим доступ ко входу и выходу нейросети, используемой в РЛС, для решения оптимизационной задачи методом дифференциальной эволюции. Так как в реальных условиях такой доступ вероятнее всего будет отсутствовать, на стороне постановщика помехи необходимо иметь ее некий эквивалент, близкий по архитектуре. Так как в работе показано, что даже небольшие отличия архитектур способны существенно снизить эффективность атаки, то выполнение этого условия само по себе представляет собой сложную задачу. Во-вторых, необходимы точные сведения о характеристиках сигнала и геометрических условиях

съемки для достоверного моделирования радиолокационного изображения прикрываемого объекта. Это изображение необходимо для подачи на вход эквивалента атакуемой сети для определения координат размещения ЛТО. И в-третьих, создание качественной (сфокусированной) ЛТО на радиолокационном изображении в точке с заданными координатами также требует точных сведений о взаимном пространственном расположении носителя РЛС и ретранслятора. При высоком разрешении РЛС ошибки позиционирования ЛТО даже в единицы метров соответствуют ее смещению до единиц или десятков пикселей на изображении. Как показано в работе, смещение даже на три-пять пикселей (то есть около одного-двух метров в случае изображений набора MSTAR), способно существенно снизить эффект от такой атаки.

3. Таким образом, требования к уровню информационного обеспечения средства противодействия, выполнение которых необходимо для обеспечения степени эффективности состязательных атак, оправдывающих усилия по их созданию, на сегодняшний день представляются практически нереализуемыми. Также необходимо отметить то соображение, что состязательная атака является малоэффективной против систем обработки, не использующих нейросетевые методы.

Не исключено, что какие-либо виды состязательных атак могут характеризоваться более универсальным характером, то есть быть менее чувствительными к изменению архитектуры атакуемой сети или пространственно-временному смещению возмущающего воздействия. В любом случае, наметившаяся в последние годы тенденция к все более широкому внедрению нейросетевых методов в практику

обработки РЛИ требует дальнейшего, хотя бы и исключительно теоретического на сегодняшний день, изучения вопросов формирования помех, способных реализовать эффект состязательной атаки.

Список источников

1. Коул А., Ганджу С., Казам М. Искусственный интеллект и компьютерное зрение. Реальные проекты на Python, Keras и TensorFlow. -СПб.: Питер, 2023. - 624 с.
2. Шолле Ф. Глубокое обучение на Python. - СПб: Питер, 2018. - 400 с.
3. Alzubaidi L., Zhang J., Humaidi A.J., Al-Dujaili A., Duan Y., Al-Shamma O., Santamaria J., Fadhel M.A., Al-Amidie M., Farhan L. Review Of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions // Journal of Big Data. 2021. Vol. 8, No. 53. URL: https://doi.org/10.1186/s40537-021-00444-8
4. Rawat W., Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review // Neural Computation. 2017. Vol. 29, P. 2352-2449. URL: https://doi.org/10.1162/neco_a_00990
5. Goodfellow I.J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. 2015. 11 p. URL: https://arxiv.org/pdf/1412.6572
6. Guo C., Gardner J.R., You Y., Wilson A.G., Weinberger K.Q. Simple Black-box Adversarial Attacks. 2019. 14 p. URL: https://arxiv.org/abs1905.07121
7. Уорр К. Надежность нейронных сетей. Укрепляем устойчивость ИИ к обману. - СПб.: Питер, 2021. - 272 с.

8. Zhou S., Liu C., Ye D., Zhu T., Zhou W., Yu P.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity // ACM Computing Surveys. 2022. Vol. 55, No. 8. Article 163. 39 p. URL: https://dl.acm.org/doi/10.1145/3547330

9. Akhtar N., Mian A. Threat of Adversarial Attacks on DL in Computer Vision: A Survey // IEEE Access 6. 2018. 21 p. URL: https://arxiv.org/pdf/1801.00553

10. Wang X., Li J., Kuang X., Tan Yu-An, Li J. The security of machine learning in an adversarial setting: A survey // Journal of Parallel and Distributed Computing. 2019. No. 130. P. 12-23. URL: https://doi.org/10.1016/j.jpdc.2019.03.003

11. Ding D., Zhang M., Feng F., Huang Y., Jiang E., Yang M. Black-Box Adversarial Attack on Time Series Classifcation // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. P. 7358-7368. URL: https://dl.acm.org/doi/abs/10.1609/aaai.v37i6.25896

12. Gao W., Liu Y., Zeng Y., Liu Q., Li Q. SAR Image Ship Target Detection Adversarial Attack and Defence Generalization Research // Sensors. 2023. No. 23. 12 p. URL: https://doi.org/10.3390/s23042266

13. Zhang Z., Gao X., Liu S., Peng B., Wang Y. Energy-Based Adversarial Example Detection for SAR Images // Remote Sensing. 2022. No. 14. 19 p. URL: https://doi.org/10.3390/rs14205168

14. Ефимов Е.Н., Шевгунов Т.Я. Идентификация точечных рассеивателей радиолокационных изображений с использованием нейронных сетей радиально-базисных функций // Труды МАИ. 2013. № 68. URL: https://trudymai.ru/published.php?ID=41959

15. Zhu X., Montazeri S., Ali M., Hua Yu., Wang Yu., Mou L., Shi Yi., Xu F., Bamler R. Deep Learning Meets SAR // Electrical Engineering and Systems Science. 2021. 26 p. URL: https://arxiv.org/abs/2006.10027

16. Anas H., Majdoulayne H., Chaimae A., Nabil S.M. Deep Learning for SAR Image Classification // Intelligent Systems and Applications, 2020. P. 890-898. URL: https://doi.org/10.1007/978-3-030-29516-5_67

17. Coman C., Thaens R. A Deep Learning SAR Target Classification Experiment on MSTAR Dataset // 19th International Radar Symposium (IRS). 2018. P. 1–6. DOI: 10.23919/IRS.2018.8448048

18. Kechagias-Stamatis O., Aouf N. Automatic Target Recognition on Synthetic Aperture Radar Imagery: A Survey // Computer Science and Engineering 2020. DOI: 10.13140/RG.2.2.16595.20008

19. Du C., Zhang L. Adversarial Attack for SAR Target Recognition Based on UNet-Generative Adversarial Network // Remote Sensing. 2021. No. 13. 20 p. URL: https://doi.org/10.3390/rs13214358

20. Li H., Huang H., Chen L., Peng J., Huang H., Cui Zh., Mei X., Wu G. Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2021. Vol. 14, P. 1333-1347. DOI: 10.1109/JSTARS.2020.3038683

21. Peng B., Peng B., Yong S., Liu L. An Empirical Study of Fully Black-Box and Universal Adversarial Attack for SAR Target Recognition // Remote Sensing. 2022. No. 14 (16). URL: https://doi.org/10.3390/rs14164017

22. Хазов А.С., Ортиков М.Ю., Гусев С.Н. Методика оценивания разрешающей способности космического радиолокатора с синтезированной апертурой антенны с учетом компенсации атмосферных искажений // Труды МАИ. 2022. № 126. URL: https://trudymai.ru/published.php?ID=169001. DOI: 10.34759/trd-2022-126-15

23. Занин К.А. Разработка модели оценки пространственного разрешения космического радиолокатора синтезированной апертуры // Труды МАИ. 2017. № 96. URL: https://trudymai.ru/published.php?ID=85931

24. Купряшкин И.Ф., Лихачев В.П. Космическая радиолокационная съемка земной поверхности в условиях помех. - Воронеж: Научная книга, 2014. - 460 с.

25. Мичурин В.В., Шабалкин А.П. Аппаратура интеллектуального подавления для защиты объектов от космического радиолокационного мониторинга // Радиотехника. 2022. Т. 86, № 5. С. 28–37. DOI: 10.18127/j00338486-202205-04

26. Гусев С.Н., Сахно И.В., Хуббиев Р.В. Методика оценивания качества формирования виртуальных объектов на радиолокационных изображениях // Труды МАИ. 2019. № 104. URL: https://trudymai.ru/published.php?ID=102169

27. Su J., Vargas D.V., Sakurai K. One Pixel Attack for Fooling DNN // IEEE Transactions on Evolutionary Computation. 2019. 15 p. URL: https://arxiv.org/abs/1710.08864

28. Купряшкин И.Ф. Сравнительные результаты точности классификации радиолокационных изображений объектов набора MSTAR сверточными нейронными сетями с различными архитектурами // Журнал радиоэлектроники. 2021. № 11. DOI: 10.30898/1684-1719.2021.11.14

29. Купряшкин И.Ф., Мазин А.С. Классификация объектов военной техники с использованием сверточной нейронной сети на радиолокационных изображениях, сформированных в условиях шумовых помех // Вестник Концерна ВКО «Алмаз – Антей». 2022. № 1. С. 71–81. DOI: 10.38013/2542-0542-2022-1-71-81

30. Price K., Storn R.M. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces // Journal of Global Optimization, 1997. Vol. 11 (4), P. 341-259. URL: https://doi.org/10.1023/A:1008202821328

31. Купряшкин И.Ф. Классификация объектов военной техники с использованием сверточной нейронной сети на радиолокационных изображениях, сформированных в условиях ретрансляционных помех // Вестник Концерна ВКО «Алмаз – Антей». 2022. № 4. С. 70–79. DOI: 10.38013/2542-0542-2022-4-70-79

32. Li K., Zhang M., Xu M., Tang R., Wang L., Wang H. Ship Detection in SAR Images Based on Feature Enhancement Swin Transformer and Adjacent Feature Fusion // Remote Sensing. 2022. No. 14. P. 3186. URL: https://doi.org/10.3390/rs14133186

33. Wickramasinghe S., Parikh D., Zhang B., Kannan R., Prasanna V., Busart C. VTR: An Optimized Vision Transformer for SAR ATR Acceleration on FPGA // Computer Science. 2024. 16 p. URL: https://arxiv.org/abs/2404.04527

34. Fein-Ashley J., Ye T., Kannan R., Prasanna V., Busart C. Benchmarking Deep Learning Classifiers for SAR Automatic Target Recognition // 2023 IEEE High Performance Extreme Computing Conference (HPEC). 2023. 6 p. DOI: 10.1109/HPEC58863.2023.10363455

35. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale // Computer Science. 2020. 22 p. URL: https://arxiv.org/abs/2010.11929

References

1. Koul A., Gandzhu S., Kazam M. *Iskusstvennyi intellekt i komp'yuternoe zrenie. Real'nye proekty na Python, Keras i TensorFlow* (Practical Deep Learning for Cloud, Mobile, and Edge). Saint Petersburg: Piter Publ., 2023. 624 p.

2. Sholle F. *Glubokoe obuchenie na Python* (Deap Learning with Python). Saint Petersburg: Piter Publ., 2018. 400 p.

3. Alzubaidi L., Zhang J., Humaidi A.J., Al-Dujaili A., Duan Y., Al-Shamma O., Santamaria J., Fadhel M.A., Al-Amidie M., Farhan L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*. 2021. Vol. 8, No. 53. URL: https://doi.org/10.1186/s40537-021-00444-8

4. Rawat W., Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*. 2017. Vol. 29, P. 2352-2449. URL: https://doi.org/10.1162/neco_a_00990

5. Goodfellow I.J., Shlens J., Szegedy C. *Explaining and Harnessing Adversarial Examples*. 2015. 11 p. URL: https://arxiv.org/pdf/1412.6572

6. Guo C., Gardner J.R., You Y., Wilson A.G., Weinberger K.Q. *Simple Black-box Adversarial Attacks*. 2019. 14 p. URL: https://arxiv.org/abs1905.07121

7. Uorr K. *Nadezhnost' neironnykh setei. Ukreplyaem ustoichivost' II k obmanu* Strenghtening (Deep Neural Networks. Making AI Less Susceptible To Adversarial Trickery). Saint Petersburg: Piter Publ., 2021. 272 p.

8. Zhou S., Liu C., Ye D., Zhu T., Zhou W., Yu P.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. *ACM Computing Surveys*. 2022. Vol. 55, No. 8. Article 163. 39 p. URL: https://dl.acm.org/doi/10.1145/3547330

9. Akhtar N., Mian A. Threat of Adversarial Attacks on DL in Computer Vision: A Survey. *IEEE Access 6*. 2018. 21 p. URL: https://arxiv.org/pdf/1801.00553

10. Wang X., Li J., Kuang X., Tan Yu-An, Li J. The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*. 2019. No. 130. P. 12-23. URL: https://doi.org/10.1016/j.jpdc.2019.03.003

11. Ding D., Zhang M., Feng F., Huang Y., Jiang E., Yang M. Black-Box Adversarial Attack on Time Series Classifcation. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023. P. 7358-7368. URL: https://dl.acm.org/doi/abs/10.1609/aaai.v37i6.25896

12. Gao W., Liu Y., Zeng Y., Liu Q., Li Q. SAR Image Ship Target Detection Adversarial Attack and Defence Generalization Research. *Sensors*. 2023. No. 23. 12 p. URL: https://doi.org/10.3390/s23042266

13. Zhang Z., Gao X., Liu S., Peng B., Wang Y. Energy-Based Adversarial Example Detection for SAR Images. *Remote Sensing*. 2022. No. 14. 19 p. URL: https://doi.org/10.3390/rs14205168

14. Efimov E.N., Shevgunov T.Ya. Identification of target scatterers in radar images using radial basis function neural networks. *Trudy MAI*. 2013. No. 68. (In Russ.). URL: https://trudymai.ru/eng/published.php?ID=41959

15. Zhu X., Montazeri S., Ali M., Hua Yu., Wang Yu., Mou L., Shi Yi., Xu F., Bamler R. Deep Learning Meets SAR. *Electrical Engineering and Systems Science*. 2021. 26 p. URL: https://arxiv.org/abs/2006.10027

16. Anas H., Majdoulayne H., Chaimae A., Nabil S.M. Deep Learning for SAR Image Classification. *Intelligent Systems and Applications*, 2020. P. 890-898. URL: https://doi.org/10.1007/978-3-030-29516-5_67

17. Coman C., Thaens R. A Deep Learning SAR Target Classification Experiment on MSTAR Dataset. *19th International Radar Symposium (IRS)*. 2018. P. 1–6. DOI: 10.23919/IRS.2018.8448048

18. Kechagias-Stamatis O., Aouf N. Automatic Target Recognition on Synthetic Aperture Radar Imagery: A Survey. *Computer Science and Engineering 2020*. DOI: 10.13140/RG.2.2.16595.20008

19. Du C., Zhang L. Adversarial Attack for SAR Target Recognition Based on UNet-Generative Adversarial Network. *Remote Sensing*. 2021. No. 13. 20 p. URL: https://doi.org/10.3390/rs13214358

20. Li H., Huang H., Chen L., Peng J., Huang H., Cui Zh., Mei X., Wu G. Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2021. Vol. 14, P. 1333-1347. DOI: 10.1109/JSTARS.2020.3038683

21. Peng B., Peng B., Yong S., Liu L. An Empirical Study of Fully Black-Box and Universal Adversarial Attack for SAR Target Recognition. *Remote Sensing*. 2022. No. 14 (16). URL: https://doi.org/10.3390/rs14164017

22. Khazov A.S., Ortikov M.Yu., Gusev S.N. A method for estimating the resolution of a space radar with a synthesized antenna aperture, taking into account the compensation of atmospheric distortions. *Trudy MAI*. 2022. No. 126. (In Russ.). URL: https://trudymai.ru/eng/published.php?ID=169001. DOI: 10.34759/trd-2022-126-15

23. Zanin K.A. Developing of a model of spacial resolution evaluation of a synthesized aperture space radar. *Trudy MAI*. 2017. No. 96. (In Russ.). URL: https://trudymai.ru/eng/published.php?ID=85931

24. Kupryashkin I.F., Likhachev V.P. *Kosmicheskaya radiolokatsionnaya s"emka zemnoi poverkhnosti v usloviyakh pomekh* (Space radar imaging of the earth's surface under interference conditions). Voronez: Nauchnaya kniga Publ., 2014. 460 p.

25. Michurin V.V., Shabalkin A.P. Intelligent suppression equipment for protecting objects from space radar monitoring. *Radiotekhnika*. 2022. Vol. 86, No. 5. P. 28–37. (In Russ.). DOI: 10.18127/j00338486-202205-04

26. Gusev S.N., Sakhno I.V., Khubbiev R.V. Evaluation technique for virtual objects on radar images formation quality. *Trudy MAI.* 2019. No. 104. (In Russ.). URL: https://trudymai.ru/eng/published.php?ID=102169

27. Su J., Vargas D.V., Sakurai K. One Pixel Attack for Fooling DNN. *IEEE Transactions on Evolutionary Computation*. 2019. 15 p. URL: https://arxiv.org/abs/1710.08864

28. Kupryashkin I.F. Comparative results of the classification accuracy of radar images of objects from the MSTAR set by convolutional neural networks with different architectures. *Zhurnal radioelektroniki*. 2021. No. 11. (In Russ.). DOI: 10.30898/1684-1719.2021.11.14

29. Kupryashkin I.F., Mazin A.S. Classification of military equipment objects using a convolutional neural network on radar images generated in noise interference conditions. *Vestnik Kontserna VKO «Almaz – Antei»*. 2022. No. 1. P. 71–81. (In Russ.). DOI: 10.38013/2542-0542-2022-1-71-81

30. Price K., Storn R.M. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, 1997. Vol. 11 (4), P. 341-259. URL: https://doi.org/10.1023/A:1008202821328

31. Kupryashkin I.F. Classification of military equipment objects using a convolutional neural network on radar images generated under relay interference. *Vestnik Kontserna VKO «Almaz – Antei»*. 2022. No. 4. P. 70–79. (In Russ.). DOI: 10.38013/2542-0542-2022-4-70-79

32. Li K., Zhang M., Xu M., Tang R., Wang L., Wang H. Ship Detection in SAR Images Based on Feature Enhancement Swin Transformer and Adjacent Feature Fusion. *Remote Sensing*. 2022. No. 14. P. 3186. URL: https://doi.org/10.3390/rs14133186

33. Wickramasinghe S., Parikh D., Zhang B., Kannan R., Prasanna V., Busart C. VTR: An Optimized Vision Transformer for SAR ATR Acceleration on FPGA. *Computer Science*. 2024. 16 p. URL: https://arxiv.org/abs/2404.04527

34. Fein-Ashley J., Ye T., Kannan R., Prasanna V., Busart C. Benchmarking Deep Learning Classifiers for SAR Automatic Target Recognition. *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. 2023. 6 p. DOI: 10.1109/HPEC58863.2023.10363455

35. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Computer Science*. 2020. 22 p. URL: https://arxiv.org/abs/2010.11929

Статья поступила в редакцию 01.04.2025
Одобрена после рецензирования 05.04.2025
Принята к публикации 25.08.2025
The article was submitted on 01.04.2025; approved after reviewing on 05.04.2025; accepted for publication on 25.08.2025