DOI: 10.15514/ISPRAS-2025-37(3)-5

Анализ и разработка методов очищения для защит метрик качества изображений

А.Е. Гущин, ORCID: 0000-0002-4055-7394 <alexanterg@gmail.com>
А.В. Анциферова, ORCID: 0000-0002-1272-5135 <aantsiferova@graphics.cs.msu.ru>
Д.С. Ватолин, ORCID: 0000-0002-8893-9340 <dmitriy@graphics.cs.msu.ru>

Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.
Исследовательский центр доверенного искусственного интеллекта ИСП РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

Аннотация. В последнее время начали исследовать область состязательных атак на метрики качества изображений, в то время как область защиты от них остается малоизученной. В данном исследовании мы стремимся охватить эту область и проверить возможность переноса защиты от атак с классификаторов изображений на методы оценки качества изображений. В этой работе мы применили несколько широко распространенных атак на модели оценки качества изображений и проверили успешность защиты от них. Методологии очистки охватывают различные техники предварительной обработки, включая геометрические преобразования, сжатие, очищение от шума и современные методы на основе нейронных сетей. Кроме того, мы рассматриваем проблему оценки эффективности методов защит, предлагая способы оценки визуального качества выходных данных и успешности нейтрализации атак. Мы тестируем защиту от атак на три метрики IQA – Linearity, MetaIQA и SPAQ.

Ключевые слова: состязательные защиты; состязательное очищение данных; оценка качества изображений.

Для цитирования: Гущин А.Е., Анциферова А.В., Ватолин Д.С. Анализ и разработка методов очищения для защит метрик качества изображений. Труды ИСП РАН, том 37, вып. 3, 2025 г., стр. 69– 84. DOI: 10.15514/ISPRAS–2025–37(3)–5.

Adversarial Purification for No-Reference Image-Quality Metrics: Applicability Study and New Methods

A.E. Gushchin, ORCID: 0000-0002-4055-7394 <alexanterg@gmail.com>
A.V. Antsiferova, ORCID: 0000-0002-1272-5135 <aantsiferova@graphics.cs.msu.ru>
D.S. Vatolin, ORCID: 0000-0002-8893-9340 <dmitriy@graphics.cs.msu.ru>

Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.
Research Centre for Trusted Artificial Intelligence of ISP RAS,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

Abstract. Recently, the area of adversarial attacks on image quality metrics has begun to be explored, whereas the area of defences remains under-researched. In this study, we aim to cover that case and check the transferability of adversarial purification defences from image classifiers to IQA methods. In this paper, we apply several widespread attacks on IQA models and examine the success of the defences against them. The purification methodologies covered different preprocessing techniques, including geometrical transformations, compression, denoising, and modern neural network-based methods. Also, we address the challenge of assessing the efficacy of a defensive methodology by proposing ways to estimate output visual quality and the success of neutralizing attacks. We test defences against attacks on three IQA metrics – Linearity, MetaIQA and SPAQ.

Keywords: adversarial attacks; adversarial purification; image quality assessment.

For citation: Gushchin A.E., Antsiferova A.V., Vatolin D.S., Adversarial purification for no-reference image-quality metrics: applicability study and new methods. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025. pp. 69-84 (in Russian). DOI: 10.15514/ISPRAS-2025-37(3)-5.

1. Введение

Метрики оценки качества изображений (IQA), основанные на машинном обучении, широко используются для разработки и оценки алгоритмов обработки изображений и видео. По сравнению с традиционными метриками, основанными на попиксельного или структурного сходства [1], метрики, основанные на машинном обучении, показывают более высокую корреляцию с субъективным качеством. Однако ряд исследований показал, что метрики качества изображений, основанные на машинном обучении, уязвимы к состязательным возмущениям, что потенциально может привести к многочисленным негативным последствиям. Во-первых, метрики IQA обычно используются как часть методологии оценки открытых бенчмарков для воспроизводимости. Участники могут использовать атаки на метрики с целью получения более высоких позиций в таблице лидеров. Обман бенчмарка - веская причина, когда результаты бенчмарка влияют на инвестиции в проект: например, разработка видеокодека требует огромных ресурсов, и обман метрики упрощает победу в открытых сравнениях. В современных видеокодеках уже есть режимы оптимизации, ориентированные на метрику (например, в кодеке Google *libaom* есть опция --tune-vmaf). Такая оптимизация может не повысить качество восприятия [2], поэтому использование нестабильных IQA-метрик в сетях доставки контента и потоковых сервисах может привести к ухудшению качества просмотра. Поскольку метрики IQA также используются для управления потоковым видео, увеличение показателя метрики может привести к увеличению битрейта после транскодирования, что негативно сказывается на канале трафика. Кроме того, метрики IQA используются в задачах компьютерного зрения, например, в медицинской сфере. Оптимизация, например, метода обработки магнитно-резонансной томографии с использованием неустойчивой метрики IQA внутри функции потерь модели может привести к некорректным результатам медицинских заключений.

Таким образом, область исследования устойчивости IQA к состязательным надбавкам начала развиваться недавно. Она не так хорошо изучена, как устойчивость методов классификации или сегментирования изображений. Для метрик качества изображений до сих пор проводились только эмпирические эксперименты с атаками противника, и никаких доказательных результатов получено не было. Среди существующих атак, предложенных для оценки устойчивости метрик, есть несколько методов, основанных на градиентной оптимизации ([3, 4, 5]), перцептивно-ориентированной маскировке [6, 7].

Механизмы защиты метрик качества изображений и видео не были тщательно изучены. Это объясняется двумя причинами: во-первых, область атак на метрики IQA находится на ранней стадии своего развития; во-вторых, постановка задачи защиты метрик IQA является более сложной, чем, например, классификация изображений. Семантика изображения остается неизменной после добавления к нему состязательной надбавки, поэтому истинная метка атакованного изображения по-прежнему известна. Однако его качество меняется, и метод защиты должен обеспечить правильную оценку исходного качества.

Насколько нам известно, только две метрики были позиционированы как более надежные версии оригинальных реализаций. Первая – VMAF [7] от Netflix, которая повышает устойчивость VMAF к улучшению изображения путем обрезания оценок базовых SVM-функций. R-LPIPS [9] и E-LPIPS [10] – это надежная версия метрики LPIPS, созданная с помощью дообучения и ансамблирования.

Нашей целью было изучить эмпирические средства защиты, полезные, когда атакующий не знает о механизме защиты, но мы также показали их эффективность против адаптивных атак. Наше исследование сосредоточено на изучении и улучшении устойчивости только метрик, не требующих исходного изображения (NR метрик) по нескольким причинам. Во-первых, эти метрики легче атаковать; они имеют больше практических применений из-за отсутствия доступа к исходному изображению (например, генерация изображений и видео, потоковое видео и т. д.). NR метрики могут быть интегрированы в функцию потерь в различных задачах компьютерного зрения. Наконец, приведенные выше примеры устойчивых метрик являются метриками, требующими исходное изображение (FR метрики); для NR метрик не было предложено устойчивых версий, и было проведено лишь несколько экспериментов.

Мы не рассматривали методы детектирования атак в этой работе, потому что они могут испытывать трудности с новыми или незаметными атаками. Методы очистки могут адаптироваться к более широкому спектру искажений и исправлять их. Также детектирование не предлагает способа правильной дальнейшей обработки атакованных изображений.

Korhonen и др. [6] показали, что состязательное дообучение помогает улучшить эмпирическую стойкость. Обрезка и изменение размеров изображений эффективно противостоят универсальным состязательным возмущениям в работе Shumitskaya и др [3]. На рис. 1 показан пример атаки на метрику Linearity и различные методы защиты от этой атаки.

Надежные метрики необходимы для разработки современных методов обработки и сжатия изображений. Такие метрики приведут к разработке надежных бенчмарков и позволят исследователям использовать метрики в составе функции потерь для обучения методов обработки изображений. Наиболее популярными методами защиты нейронных сетей являются состязательное дообучение, очистка данных и детектирование атак. Данная статья посвящена эмпирическим механизмам защиты как первому шагу к улучшению устойчивости метрик. Несмотря на то, что состязательное дообучение может эффективно препятствовать атакам, оно также снижает корреляции метрик с субъективными оценками. В данной работе изучается применимость очистки изображений для защиты метрик IQA и предлагаются новые методы очистки.

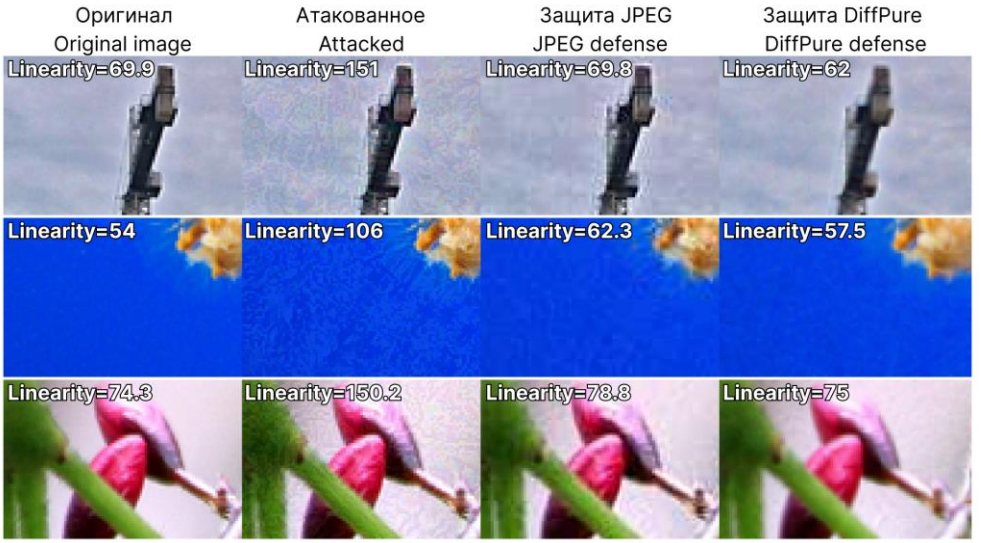

Рис. 1. Оригинальное изображение (первый столбец), атакованное изображение (второй столбец), пример работы двух методов защит – JPEG и DiffPure (третий и четвертый столбцы, соответственно).

Fig. 1. Original image (first column), image after the adversarial attack (AMI-FGSM, second columns), and two defence techniques applied to the adversarial image (third and fourth columns).

2. Обзор литературы

2.1 Состязательные атаки на метрики оценки качества

Szegedy и другие [11] показали, что нейронные сети уязвимы к специально сгенерированным входным данным, что приводит к неправильному выводу модели. Впоследствии это стало известно как состязательные атаки. Было предложено большое количество методов атаки, особенно для задачи классификации изображений. Большинство из этих атак [12, 13, 14, 15] являются аддитивными, поскольку к исходному изображению добавляется дополнительный шум, создаваемый противником. Эти методы решают оптимизационную задачу в пространстве пикселей с некоторыми ограничениями на величину возмущения в терминах нормы lp. Поскольку было показано, что lp неэффективна для аппроксимации визуального качества, были предложены состязательные атаки, сохраняющие качество восприятия, и методы преобразования цвета [16]. Несмотря на то, что большинство существующих атак создано для моделей компьютерного зрения, существуют атаки, специально разработанные для метрик IQA. AMI-FGSM [17] – это атака на метрики NR, которая является модификацией FGSM [13]. Shumitskaya и др. [3] предложили обучение универсального возмущения на наборе изображений. Ghildyal & Liu [18] обратили внимание на уязвимость FR метрик к атакам. Они адаптировали FGSM и PGD атаки для создания примеров атак на метрики IQA. Korhonen и др. [6] предложили итеративную атаку на NR метрики, которая улучшает визуальное качество атакованных изображений с помощью фильтра Собеля.

2.2 Состязательные защиты

Как и для методов атаки, для задачи классификации изображений было предложено множество способов защиты. Модификация процесса обучения или применения модели может повысить устойчивость к атакам. Одним из наиболее популярных методов защиты моделей от атак противника является состязательное дообучение [13, 15]. Оно предполагает

динамическое добавление в обучающий набор данных изображений, сгенерированных противником. Такой подход не только требует модификации процедуры обучения, но и увеличивает вычислительные затраты, так как на каждой итерации необходимо выполнить дополнительно один или даже несколько прямых и обратных проходов для генерации состязательных примеров. Кроме того, обучение на состязательных примерах приводит к снижению производительности на чистых входных данных. Аналогичный эффект от обучения с использованием состязательных примеров был продемонстрирован в работе Korhonen & You [6] для моделей IQA.

Другой тип защиты – предварительная обработка входных данных с помощью вспомогательных преобразований перед основной моделью, также называемая очисткой. Этот тип защиты требует модификации выводов модели. Graece и другие [19] предложили метод очистки, основанный на простых преобразованиях изображений, таких как обрезка и размытие. Guo и другие [20] предложили защищаться от атак противника путем усреднения предсказаний нескольких случайных участков изображений во время оценки. Этот подход направлен на изменение пространственного расположения неблагоприятных возмущений, которые не инвариантны к таким преобразованиям. В ряде работ [21, 22, 20] в качестве метода предотвращения атак противника предложено сжатие JPEG. Xu и др. [23] предложили подход, основанный на сжатии признаков внутри модели для уменьшения размеров входного изображения, чтобы снизить возможности злоумышленника по проведению атаки. Основная идея заключается в удалении ненужных признаков путем уменьшения битовой глубины цвета или применения пространственного сглаживания. Guo и другие изучали удаление отдельных пикселей с минимизацией общей вариации [24]. Этот подход предполагает выбор небольшого набора пикселей и восстановление изображения без состязательного шума. Meng и Chen в своей защите MagNet [25] использовали автокодировщик для перемещения состязательных примеров ближе к множеству чистых примеров. Однако этот метод предназначен для сценариев, в которых злоумышленник не может получить доступ к параметрам модели, и, следовательно, не подходит для атак на основе градиента, которые представляют собой значительную угрозу. С ростом популярности генеративных моделей они также стали использоваться для очистки от атак. В частности, Samangouei и другие [26] предложили использовать генеративные состязательные сети (GAN), которые удаляют состязательные возмущения из изображений. Недавно Nie и другие [27] предложили использовать все более популярные диффузионные модели для удаления неблагоприятных возмущений из изображений. Идея заключается в том, что в процессе прямой диффузии вредные возмущения должны смешиваться с шумом, а затем обратный процесс удаляет и шум, и возмущения, в результате чего получается чистое изображение.

Не все методы очистки классификаторов подходят для регрессионных моделей оценки качества, поскольку качество изображения не является решающим фактором в работе классификатора. В задаче IQA важно не только вернуть исходные оценки, но и не вносить существенных изменений в изображение. Для этих моделей было предложено всего несколько методов защиты. Korhonen и др. [6] использовали гауссово размытие и двустороннюю фильтрацию для защиты от атак. Авторы делают вывод, что методы предварительной обработки заслуживают детального изучения в будущем, что и представлено в нашей работе.

3. Методология

3.1 Постановка задачи

В нашей работе рассматриваются атаки, направленные на повышение прогнозируемых оценок качества. Эти атаки были предложены в последних работах и имеют большую практическую применимость [6, 3]. Атаки, направленные на снижение прогнозируемого

качества изображения, выглядят аналогично, за исключением знака вектора, обозначающего направление атаки. Таким образом, данный подход не нарушает общности исследования.

Для заданной NR метрики качества изображения f и входного изображения $x \in R^{H\times W\times 3}$, атака противника, направленная на увеличение прогнозируемой оценки качества, может быть математически описана следующим образом:

$$maxf\big(att(x)\big), \text{такое что} dist\big(x, att(x)\big) \leq \epsilon \quad (1)$$

где $att: R^{H\times W\times 3} \to R^{H\times W\times 3}$ – алгоритм атаки противника, применяемый к x, $dist(\cdot)$-некоторая метрика расстояния, а ε – ограничение на расстояние. Атака считается успешной, если она увеличивает предсказанную оценку, но не улучшает визуальное качество. Таким образом, изменения в изображении должны быть незаметны. Состязательная защита – это процесс, позволяющий уменьшить улучшение качества, вызванное атакой:

$$min\left(\left|f\left(g\big(att(x)\big)\right) - f(x)\right| + \lambda dist\left(x, g\big(att(x)\big)\right)\right) \quad (2)$$

где $g: R^{H\times W\times 3} \to R^{H\times W\times 3}$ – метод защиты. Защита направлена на то, чтобы вернуть атакованное изображение к оригиналу без уменьшения корреляции с субъективным качеством.

3.2 Состязательные атаки

Мы выбрали 10 методов атак для оценки методов защит. Большинство из них первоначально были представлены как атаки на модели классификации изображений; таким образом, мы адаптировали их к задаче IQA, заменив функцию потерь, которая увеличивает оценку качества:

$$L(\theta, x) = 1 - \frac{f_\theta(x)}{max(f_\theta) - min(f_\theta)} \quad (3)$$

(3) где $max(f_\theta)$ и $m \in (f_\theta)$ представляют собой наибольшую и наименьшую субъективные оценки соответственно, полученные метрикой на наборе данных NIPS 2017: Adversarial Learning Development Set [28].

Ниже мы опишем атаки, используемые в нашей работе. Первая группа атак – это метод быстрого знакового градиента (FGSM) и его вариации. FGSM [25] делает один шаг в направлении, противоположном градиенту, чтобы минимизировать объективную функцию (ур. 3). Его модификация I-FGSM [14] делает несколько итераций маленьких шагов в направлении градиента. MI-FGSM [29] дополнительно использует импульс при оптимизации. AMI-FGSM [30] был представлен как атака на модели IQA и представляет собой версию MI-FGSM, которая регулирует допустимую величину атаки путем ограничения качества восприятия, вычисляемого с помощью NR метрики NIQE [31].

В других выбранных методах используются различные техники, чтобы атака оставалась невидимой для человеческого глаза. Korhonen и др. [6] – метод генерации неблагоприятных изображений для оценки качества NR с использованием структурной информации для концентрации возмущений в текстурных областях. SSAH [12] определяет местоположение возмущений в высокочастотных компонентах, используя низкочастотные ограничения. MADC [32] – это метод сравнения метрик качества изображения. Он синтезирует пару изображений, чтобы максимизировать результат одной метрики IQA при неизменном значении другой метрики. Мы используем MADC в качестве атаки на метрику IQA, двигаясь в направлении метрики при фиксированных значениях MSE, как это было предложено в работе Antsiferova и др. [5]. предложили добавить FR метрику в качестве дополнительного элемента оптимизационной функции для создания атаки на метрику NR IQA. Мы использовали три версии этой атаки, где для сохранения визуального качества

использовались метрики SSIM [34], LPIPS [35] и DISTS [36]. Описанные выше атаки являются ограниченными, поскольку возмущения, добавляемые к исходному изображению, ограничены нормой lp или другими метриками качества. Кроме того, мы используем неограниченную атаку на цветовой фильтр AdvCF [16], которая изменяет цвета изображения путем оптимизации простого цветового фильтра, широко используемого в популярных программах для редактирования фотографий.

3.3 Методы очищения данных

Мы исследуем несколько стандартных методов предварительной обработки для очистки изображений от аддитивного состязательного шума. Первый метод, который мы выбрали, – это техника изменения размера, которая меняет разрешение входных изображений на меньшее, а затем возвращает их к исходному разрешению. Мы провели эксперименты с различными режимами интерполяции. В качестве базового был выбран билинейный режим. Следуя примеру Guo и др. [20], мы случайным образом обрезали и изменяли размер изображений до исходного. Мы также рассмотрели защиту JPEG, которая широко используется для защиты классификации изображений. Кроме того, в наш список методов вошли пространственные преобразования, такие как переворот (зеркальное отображение изображения) и случайное вращение. По аналогии с экспериментами по IQA, проведенными Korhonen и др. [6], мы применили гауссово размытие и двустороннюю фильтрацию. Гауссовый фильтр сглаживает все изображение без учета деталей. Билатеральный фильтр, напротив, нелинейный, учитывает значения интенсивности соседних пикселей, тем самым сохраняя края при сглаживании. Кроме того, мы оценили медианный фильтр, который заменяет каждый пиксель медианой соседних пикселей. Также мы добавили композицию из двух методов: нерезкое маскирование и гауссово размытие. Суть метода заключается в следующем: сначала мы обесцвечиваем изображения и отфильтровываем высокочастотный шум, а затем повышаем резкость краев объектов. Таким образом, на первом этапе происходит фильтрация неблагоприятных возмущений, а на втором – восстановление качества очищенного изображения. Диффузионная модель DiffPure [27] показала свою эффективность в очистке состязательных изображений. Небольшое количество шума вносится в атакованное изображение с помощью прямой прохода модели. Процесс диффузии продолжается до тех пор, пока не будет достигнут оптимально рассчитанный временной шаг. Затем модель делает обратный проход чтобы восстановить очищенное изображение. Цель состоит в том, чтобы возмущения постепенно слились с шумом, в результате чего добавленный гауссовский шум станет доминирующим. В результате обратный процесс удаляет добавленный шум и неблагоприятные возмущения, что приводит к получению очищенного изображения.

Из-за эффекта размытия, создаваемого DiffPure, и с учетом того, что метрики качества снижают свои значения из-за размытия, мы реализовали две дополнительных модификации метода DiffPure для решения этой проблемы. Первый называется DiffPure+Edge и основан на смешивании атакованных и очищенных изображений. Мы выделяем края изображения с помощью фильтра Собеля, а затем заменяем краевые пиксели на краевые пиксели из атакованного изображения. Это позволяет повысить резкость изображения. Второй вариант (DiffPure+Unsharp) аналогичен методу Unsharp. Сначала к атакованному изображению применяется DiffPure для удаления всех нежелательных шумов, а затем нерезкое маскирование повышает резкость изображения, улучшая его визуальное качество.

Кроме того, мы исследуем популярные методы восстановления изображений, такие как MPRNet [37] и RealESRGAN [38], для очистки изображений от состязательных надбавок. MPRNet – это трехступенчатая сверточная нейронная сеть, предназначенная для решения трех задач восстановления изображений, таких как размытие, удаления эффекта дождя и удаления шума. Первые два этапа сети используют архитектуру кодировщик-декодировщик для извлечения контекстной информации на нескольких разрешениях. На последнем этапе изображение обрабатывается в исходном разрешении, чтобы сохранить мелкие детали.

Ключевыми особенностями MPRNet являются модули внимания, расположенные между этапами, и слияние признаков между этими этапами, которое обеспечивает эффективную передачу информации от ранних этапов к поздним. Real-ESRGAN [38] – это модель восстановления, основанная на генеративной состязательной сети, обученной на синтетических данных. Авторы использовали глубокую сеть, состоящую из нескольких остаточных линейных блоков, для выполнения задачи увеличения разрешения. Мы установили масштабный коэффициент 1, чтобы сохранить исходный размер изображения.

Ранее все средства защиты были разработаны для снижения уровня состязательной надбавки в изображениях, что делает их неэффективными против цветных атак, таких как AdvCF [16]. Мы предлагаем новый метод защиты, основанный на компактной полностью сверточной нейронной сети (FCN-фильтр). Наша модель состоит из трех сверточных слоев, которые применяют 64, 16 и 3 фильтра и сохраняют исходные размеры изображения. Набор данных для обучения - 200 изображений с NIPS 2017, атакованных AdvCf, из которых обучающая часть составляет 80 %, а валидационная – оставшиеся 20 %. Мы обучаем его с помощью оптимизатора Адама в течение 200 эпох с коэффициентом обучения 1e-3 для восстановления оригинального изображения из атакованного, оптимизируя среднюю квадратичную ошибку (MSE). Мы сохраняем модель, показавшую наилучшие результаты на валидационном множестве по показателю SSIM между очищенными и оригинальными изображениями.

3.4 Набор атакованных данных

Мы создали набор данных с атакованными изображениями, чтобы проанализировать эффективность защиты от используемых атак. В качестве исходных изображений использовались изображения из набора данных NIPS 2017: Adversarial Learning Development Set [28], которые подвергались дальнейшим атакам. Этот набор данных ранее использовался для решения широкого круга задач компьютерного зрения, включая атаки на метрики IQA [39]. Атака на метод Linearity проводилась на каждом изображении с помощью всех методов атаки из раздела 3.2 с использованием параметров по умолчанию. Таким образом, из каждого чистого изображения мы получили 10 атакованных изображений. Примеры атакованных изображений представлены на рис. 2. Таблица всех параметров, которые использовались в атаках, представлена в Дополнительных материалах.

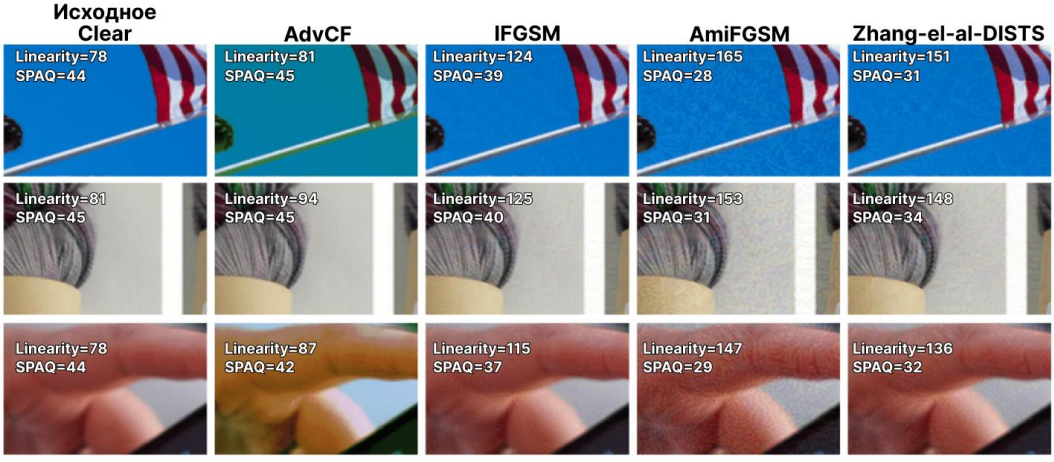

Рис. 2. Примеры атакованных изображений. Первая колонка показывает исходное изображение, следующие показывают атакованные изображения после AdvCF, IFGSM, AmiFGSM, Zhang атак, соответственно.

Fig. 2. Examples of attacks. The first column shows the original (clean) image, the following columns display the corresponding image after the attack.

В качестве атакуемых моделей для атак противника мы использовали три NR метрики качества изображения – Linearity [40], MetaIQA [41] и SPAQ [42]. Авторы Linearity создали собственную функцию потерь, которая сходится примерно в 10 раз быстрее, чем популярные функции потерь MAE и MSE. Linearity, MetaIQA и SPAQ были выбраны в качестве атакуемой метрики, поскольку они демонстрируют высокую производительность (корреляция с субъективным качеством и скоростью) и среднюю устойчивость к атакам противника:

- Корреляция с субъективными оценками и высокая скорость. В MSU No-Reference Video Quality Metrics Benchmark [43] показывает хорошую производительность среди всех метрик качества изображения без ссылок. Эти метрики также отличаются высокой скоростью вычисления по сравнению с аналогами [44].
- Средняя устойчивость. Для нашего исследования нам нужна была целевая метрика IQA со средней устойчивостью к атакам противника. Атаки на высокоустойчивые метрики приведут к получению изображений, которые будут сильно отличаться от чистых, или же атака будет малоэффективна для повышения значений метрики. Таким образом, практическая польза от таких атак незначительна. При использовании метрик с низкой устойчивостью атаки будут визуально незаметны. Такие атаки можно нейтрализовать, применив гауссово размытие с малым размером ядра. Согласно [39], метрики Linearity, MetaIQA и SPAQ демонстрируют среднюю устойчивость по сравнению с другими метриками, что делает ее подходящей целью для нашей работы.

3.5 Детали реализации

Мы использовали открытый исходный код метрик без дополнительного предварительного обучения и выбрали параметры по умолчанию. Гиперпараметры атак также были выставлены по умолчанию. Для обеспечения полной воспроизводимости результатов были использованы инструменты GITLAB CI/CD. Расчеты проводились на компьютере со следующими характеристиками: NVIDIA A100-PCIE-40GB, 32-ядерный процессор Intel Xeon Processor (Icelake) @ 2,90 ГГц. Все расчеты заняли в общей сложности около 50 часов работы GPU.

3.6 Метрики оценивания эффективности

Мы используем две группы метрик для оценки успешности защиты и качества восприятия очищенного изображения для сравнения методов защиты.

Метрики качества оценивают эффективность защиты при обработке изображения с точки зрения визуального качества. Мы выбрали метрики качества изображения PSNR и SSIM из-за их большей устойчивости к атакам противника по сравнению с метриками No-Reference. Кроме того, недифференцируемость этих метрик ставит перед злоумышленниками более сложную задачу. Чтобы объединить эти две метрики в одну оценку, мы складываем их с линейным коэффициентом:

$$Qualityscore = \frac{PSNR}{100} + \frac{SSIM}{2} \quad (4)$$

Эти коэффициенты были выбраны по следующим причинам: для SSIM максимальное значение равно 1. PSNR со значением более 50 означает, что сравниваемые изображения полностью неразличимы, хотя PSNR не имеет максимального значения.

Метрика прироста отражает эффективность защиты по показателям значения метрики. Мы рассчитываем относительный выигрыш для атакованных и очищенных (после последовательного применения атаки и защиты) изображений следующим образом:

$$Gainscore = \frac{|Clear - Distorted|}{Clear} \quad (5)$$

где Distorted – это либо атакованное, либо очищенное изображение, соответствующее чистому изображению. Мы также рассчитали коэффициент корреляции Спирмена (SROCC) между значениями метрики на чистых и очищенных изображениях.

4. Результаты

Мы выделяем три основные характеристики, которыми должны обладать средства защиты: нейтрализация эффектов атак противника, восстановление исходного качества изображения и сохранение корреляций значений атакуемой модели. Для оценки этих характеристик мы сравниваем методы очистки по показателям относительного выигрыша, метрикам PSNR/SSIM и коэффициенту корреляции Спирмена (SROCC) на наборе данных с атаки противника, описанном в разделе 3.4.

Общая эффективность. Табл. 1 показывает усредненные результаты для чистых и атакованных изображений вместе. По показателю качества выходного изображения лидируют методы Upscale, MPRNet и Unsharp. Тем не менее методы DiffPure и DiffPure+Unsharp демонстрируют сопоставимую производительность, отставая от лидера менее чем на 4% как по метрике PSNR, так и по метрике SSIM. Следующие методы лучше всего нейтрализуют последствия атаки DiffPure+Unsharp, DiffPure и изменение размера. Кроме того, DiffPure и DiffPure+Unsharp, наряду с билатеральной фильтрацией, лучше всего сохраняют корреляции между значениями метрик. Метод защиты, заключающийся только в переворачивании изображения по обеим осям, показывает наилучший показатель прироста. Таким образом, этот метод может использоваться для получения значения модели на исходном изображении, при этом по показателю качества он является наихудшим среди рассмотренных. Однако, любую атаку легко адаптировать против такой защиты. RealESRGAN демонстрирует относительно хороший показатель качества, но не справляется с последствиями атак противника и не сохраняет стабильность значений моделей в терминах SROCC. Метод хорошо работает в низкочастотных областях, но создает визуально неприятные артефакты в областях с высокой текстурой. Метод DiffPure+Edge получил один из лучших показателей качества, но показал низкий коэффициент SROCC и показатель прироста. Причиной этого может быть сильный акцент на краях при расчете значений моделей качества.

Подводя итог, можно сказать, что метод DiffPure является наиболее предпочтительным с точки зрения всех трех сравниваемых характеристик. Кроме того, добавление нерезкого маскирования после применения DiffPure несколько повышает их все. Основным недостатком этой техники является ее высокая вычислительная сложность, что приводит к увеличению времени выполнения. Табл. 1 также показывает среднюю скорость вычислений (FPS) для каждого метода.

Эффективность на чистых изображений. При применении защиты обычно неизвестно заранее, производилась ли атака на это изображение, поэтому мы проверяем эффективность после применения всех защит и на чистых изображениях. Эксперименты показывают, что по показателю прироста простые преобразования, такие как поворот и кадрирование, показывают наилучшие результаты при применении к чистым изображениям. Фильтр FCN не сильно отстает, занимая третье место. Все три метода демонстрируют хороший коэффициент SROCC (выше 0,94), при этом практически не изменяя значения моделей с показателем прироста до 0,04. Однако по показателю качества поворот и обрезка находятся в самом низу рейтинга, в то время как фильтр FCN имеет высокое среднее значение SSIM, превышающее 0,9. По показателю качества лидерами являются четыре метода с очень похожими результатами – DiffPure+Edge, MPRNet, нерезкое маскирование и метод Upscale. Более подробно данные представлены в Табл. 2.

Качество изображений после применения защит. Рис. 3 (слева) иллюстрирует значения показателя прироста для изображений с различными минимальными значениями метрики SSIM. Он считался для всех очищенных изображений, для которых показатель SSIM между соответствующими исходными и атакованными изображениями не меньше порога (точки на оси X). Некоторые методы были исключены для лучшего визуального представления. Для большинства методов показатель прироста остается стабильным при любых значениях SSIM.

Некоторые методы, такие как Upscale, FCNфильтр и DiffPure+Edge, ведут себя по-разному при слабых атаках (таких, после которых метрика SSIM имеет высокие значения).

Табл. 1. Усредненные показатели эффективности на атакованных изображениях.
Table 1. Results for attacked images.

Метод	Показатель качества ↑	Показатель прироста ↓	SROCC↑	FPS↑
Билатеральный фильтр	0.594	0.151	**0.731**	6.78
Зеркалирование	0.158	0.077	0.728	**36.67**
Размытие	0.625	0.138	0.471	16.65
JPEG	0.626	0.099	0.725	15.76
Медианный фильтр	0.615	0.118	0.671	16.56
MPRNet	0.653	0.179	0.611	6.33
Обрезка	0.227	0.247	0.626	29.92
Real-ESRGAN	0.636	0.360	0.562	10.15
Resize	0.626	0.099	0.725	**63.31**
Поворот	0.344	0.239	0.501	**36.18**
Нерезкое маскирование	0.649	0.133	0.513	15.96
Upscale	0.632	0.207	0.586	35.50
FCN (предложен)	0.591	0.409	0.619	35.56
DiffPure	0.628	0.093	0.706	3.08
+Edge (предложен)	**0.660**	0.195	0.480	2.98
+Unsharp (предложен)	0.632	**0.062**	**0.750**	2.90

Табл.2. Усредненные показатели эффективности на чистых изображениях.
Table 2. Results for clean images.

Метод	Показатель качества ↑	Показатель прироста ↓	PSNR↑	SSIM↑	SROCC↑
Билатеральный фильтр	0.605	0.068	29.185	0.846	0.826
Зеркалирование	0.153	0.043	9.744	0.185	0.886
Размытие	0.651	0.235	30.433	0.923	0.752
JPEG	0.641	0.168	30.585	0.899	0.849
Медианный фильтр	0.635	0.080	29.867	0.896	0.875
MPRNet	0.667	0.070	31.147	0.932	0.854
Обрезка	0.242	**0.022**	13.655	0.314	0.972
Real-ESRGAN	0.639	0.073	28.250	0.925	0.812
Resize	0.641	0.168	30.585	0.899	0.849
Поворот	0.393	0.036	22.633	0.502	0.962
Нерезкое маскирование	0.663	0.206	31.476	0.932	0.772
Upscale	0.665	0.160	31.282	0.931	0.867
FCN (предложен)	0.606	0.038	24.535	0.906	**0.976**
DiffPure	0.637	0.149	30.147	0.898	0.955
+Edge (предложен)	**0.682**	0.074	**32.170**	**0.933**	0.958
+Unsharp (предложен)	0.643	0.115	30.584	0.904	0.918

Рис. 3 (справа) показывает значения показателя прироста, основанные на минимальном значении атакуемой модели. В крайней правой части графика показатель рассчитывался

только для сильных атак (то есть значение атакуемой модели велико). В крайнем левом случае он был рассчитан для всего набора данных. Мы видим, что показатель прироста монотонно изменяется по сравнению со значением модели после атаки. Три метода (DiffPure, фильтр Гаусса и нерезкое маскирование) выделяются на фоне остальных – их показатель прироста монотонно уменьшается, показывая, что они лучше справляются с последствиями сильных атак.

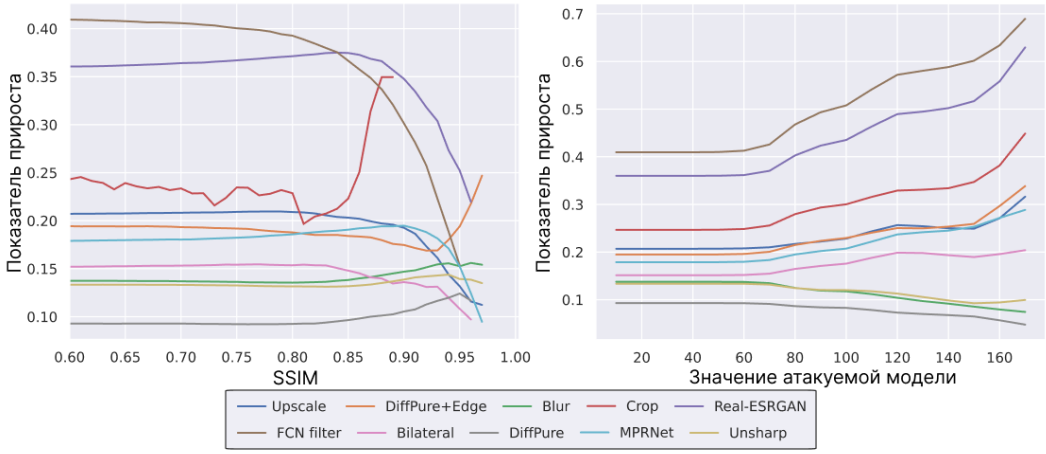

Рис. 3: Gain score в зависимости от минимальных значений SSIM (слева), значения модели после атаки (справа).
Рис. 3: Gain score as a function of minimum SSIM values (left), model's values after attack (right).

Субъективное оценивание. Мы провели субъективное исследование, оценив качество восприятия изображений, полученных с помощью методов защит на краудсорсинговой платформе Subjectify.us. Subjectify.us – это сервис для парных сравнений; он использует модель Брэдли-Терри для преобразования результатов парного голосования в оценку каждого изображения.

Для каждого набора из модели оценки качества, атаки, силы атаки и защиты было выбрано 10 случайных изображений, в результате чего получилось 10 (количество чистых изображений) * 14 (количество атак, включая изображения без атак) * 3 (количество параметров для каждой атаки) * 13 (количество методов защиты) = 5,460 изображений для оценки. Мы исключили 3 метода защиты с простыми геометрическими преобразованиями: поворот, зеркалирование и обрезка, поскольку их качество нельзя напрямую сравнивать с другими методами в субъективном исследовании.

Участникам исследования нужно было оценить пару изображений и указать какое из них имеет лучшее визуальное качество. Каждое изображение в паре имело одно исходное изображение, одну атаку и ее параметры. Различие между ними было только в примененной защите. Участники последовательно просматривали изображения из каждой пары в полноэкранном режиме. Каждый участник должен был сравнить 25 пар, две из которых имели вариант с более высоким качеством и служили проверочными вопросами. Все ответы тех, кто не смог правильно ответить на проверочные вопросы, были отброшены.

Чтобы обеспечить достоверность результатов, было получено не менее 10 голосов для каждой пары изображений.

В общей сложности мы собрали 72 425 ответов от 2 897 человек. Применив модель Брэдли-Терри к таблице парных рангов, мы получили субъективные оценки для каждого изображения. Результаты субъективного сравнения можно найти на рис. 4.

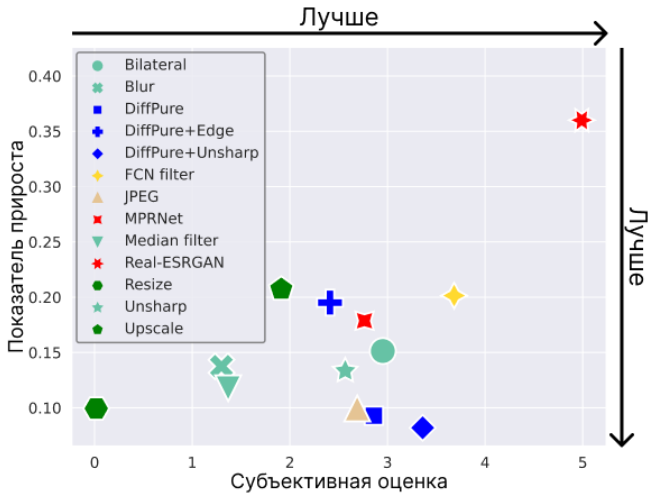

Рис. 4: Субъективные оценки качества и показатель прироста для различных защит, усредненные по всем атакам.
Рис. 4: Subjective scores and gain score for different defences averaged for all attacks.

5. Заключение

В данной работе мы исследовали эффективность методов очистки от состязательных атак в качестве защиты от атак на метрики оценки качества изображений. Мы провели обширное исследование, включающее 10 атак и 16 методов очистки, и опубликовали набор данных изображений, подвергшихся атакам.

Результаты показали, что даже простые и быстрые методы, такие как поворот или зеркалирование изображения, могут нивелировать эффекты атаки и сохранить визуальное качество очищенного изображения близким к исходному. Более сложные средства защиты могут сохранять и восстанавливать исходное качество: предложенные нами комбинации DiffPure с нерезким маскированием обеспечивают наивысший SROCC, сохраняя высокое качество очищенных изображений и Gain score. Эффективность протестированных защит схожа для различных итеративных градиентных атак, но отличается для неограниченных атак типа AdvCf. Такие атаки могут быть нейтрализованы путем обучения нейронной сети на состязательных примерах. В нашем исследовании мы предложили фильтр FCN, который эффективно защищает метрики оценки качества от атаки AdvCF.

Список литературы / References

[1]. Duanmu, Z., Liu, W., Wang, Z., Wang, Z.: Quantifying visual image quality: A bayesian view. An-nual Review of Vision Science 7, 437–464 (2021).
[2]. Zvezdakova, A., Zvezdakov, S., Kulikov, D., Vatolin, D.: Hacking vmaf with video color and con-trast distortion. arXiv preprint arXiv:1907.04807 (2019).
[3]. Shumitskaya, E., Antsiferova, A., Vatolin, D.: Universal perturbation attack on differentiable no-reference image- and video-quality metrics (2022).
[4]. Shumitskaya, E., Antsiferova, A., Vatolin, D.: Towards adversarial robustness verification of no-reference image-and video-quality metrics. Computer Vision and Image Understanding 240, 103913 (2024).
[5]. Zhang, W., Li, D., Min, X., Zhai, G., Guo, G., Yang, X., Ma, K.: Perceptual attacks of no-reference image quality models with human-in-the-loop. Advances in Neural Information Processing Systems 35, 2916–2929 (2022).
[6]. Korhonen, J., You, J.: Adversarial attacks against blind image quality assessment models. Proceed-ings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications (2022), https://api.seman-ticscholar.org/CorpusID: 252546140 16 F. Author et al.

[7]. Luo, C., Lin, Q., Xie, W., Wu, B., Xie, J., Shen, L.: Frequency-driven imperceptible adversarial at-tack on semantic similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15315–15324 (June 2022).
[8]. Li, Z.: On vmaf's property in the presence of image enhancement operations (2021).
[9]. Ghazanfari, S., Garg, S., Krishnamurthy, P., Khorrami, F., Araujo, A.: R-lpips: An adversarially ro-bust perceptual similarity metric. arXiv preprint arXiv:2307.15157 (2023).
[10]. Kettunen, M., Härkönen, E., Lehtinen, J.: E-lpips: robust perceptual image similarity via random transformation ensembles. arXiv preprint arXiv:1906.03973 (2019).
[11]. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. CoRR abs/1312.6199 (2013), https://api.semanticscholar.org/CorpusID:604334 (дата обращения 12.09.2024).
[12]. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks (2017).
[13]. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015).
[14]. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world (2017).
[15]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2019).
[16]. Zhao, Z., Liu, Z., Larson, M.: Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter (2020).
[17]. Sang, Q., Zhang, H., Liu, L., Wu, X., Bovik, A.: On the generation of adversarial samples for image quality assessment. SSRN Electronic Journal (01 2022). https: //doi.org/10.2139/ssrn.4112969 (дата обращения 12.09.2024).
[18]. Ghildyal, A., Liu, F.: Attacking perceptual similarity metrics (2023).
[19]. Graese, A., Rozsa, A., Boult, T.E.: Assessing threat of adversarial examples on deep neural networks (2016).
[20]. Guo, C., Rana, M., Cisse, M., van der Maaten, L.: Countering adversarial images using input trans-formations (2018).
[21]. Das, N., Shanbhogue, M., Chen, S.T., Hohman, F., Chen, L., Kounavis, M.E., Chau, D.H.: Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression (2017).
[22]. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images (2016).
[23]. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural net-works. CoRR abs/1704.01155 (2017), http://arxiv.org/abs/ 1704.01155 (дата обращения 12.09.2024).
[24]. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60, 259–268 (1992), https://api. semanticscholar.org/CorpusID:13133466.
[25]. Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples (2017).
[26]. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models (2018).
[27]. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversar-ial purification (2022).
[28]. Nips 2017: Adversarial learning development set. https://www.kaggle.com/ datasets/google-brain/nips-2017-adversarial-learning-development-set (2017) (дата обращения 12.09.2024).
[29]. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momen-tum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018).
[30]. On the generation of adversarial examples for image quality assessment. Visual Computer (2023). https://doi.org/10.1007/s00371-023-03019-1 (дата обращения 12.09.2024).
[31]. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters 20, 209–212 (2013), https://api. semanticscholar.org/CorpusID:16892725.
[32]. Wang, Z., Simoncelli, E.: Maximum differentiation (mad) competition: A methodology for compar-ing computational models of perceptual quantities. Journal of vision 8, 8.1–13 (02 2008). https://doi.org/10.1167/8.12.8 (дата обращения 12.09.2024).
[33]. Antsiferova, A., Abud, K., Gushchin, A., Shumitskaya, E., Lavrushkin, S., Vatolin, D.: Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks (2024).
[34]. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861 (дата обращения 12.09.2024).

[35]. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pat-tern recognition. pp. 586–595 (2018).
[36]. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and tex-ture similarity. IEEE transactions on pattern analysis and machine intelligence 44(5), 2567–2581 (2020).
[37]. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progres-sive image restoration. In: CVPR (2021) Adversarial purification for no-reference image-quality met-rics 17.
[38]. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind superresolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW).
[39]. Antsiferova, A., Abud, K., Gushchin, A., Shumitskaya, E., Lavrushkin, S., Vatolin, D.: Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks (2024).
[40]. Li, D., Jiang, T., Jiang, M.: Norm-in-norm loss with faster convergence and better performance for image quality assessment. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 789–797 (2020).
[41]. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: Metaiqa: Deep meta-learning for noreference image quali-ty assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-nition. pp. 14143–14152 (2020).
[42]. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photog-raphy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3677–3686 (2020).
[43]. Antsiferova, A., Lavrushkin, S., Smirnov, M., Gushchin, A., Vatolin, D., Kulikov, D.: Video com-pression dataset and benchmark of learning-based video-quality metrics. Advances in Neural Information Processing Systems 35, 13814–13825 (2022).
[44]. https://videoprocessing.ai/benchmarks/video-quality-metrics_nrm.html (дата обращения 12.09.2024).

Информация об авторах / Information about authors

Александр Евгеньевич ГУЩИН получил степень магистра по прикладной математике и информатике в Московском государственном университете имени М. В. Ломоносова в 2024 году. В настоящее время учится в аспирантуре и ведет исследования в лаборатории компьютерной графики и мультимедиа, а также в Центре доверенного искусственного интеллекта ИСП РАН. В область его научных интересов входят методы оценивания качества видео и изображений, а также исследование устойчивости нейросетевых моделей и методов их защиты.

Aleksandr Evgenevich GUSHCHIN received his master degree in computer science from the Moscow State University in 2024. He is currently a postgraduate student at the MSU Graphics & Media Lab, and a researcher at the Research Centre for Trusted AI of ISP RAS. His research interests involve image and video processing, quality assessment, and machine learning. He is also a key contributor to the project analyzing video quality assessment methods, including their robustness to adversarial attacks.

Анастасия Всеволодовна АНЦИФЕРОВА аспирантка ВМК МГУ. Окончила магистратуру ВМК МГУ по специальности анализ больших данных в 2018 году. В настоящее время она является аспирантом и участником видеогруппы лаборатории графики и мультимедиа МГУ, ведет исследования в Центре доверенного искусственного интеллекта ИСП РАН. Сфера ее научных интересов включает анализ и оптимизацию видеокодеков, оценку субъективного качества стереоскопического видео. Анастасия является одним из организаторов проекта международного сравнения видеокодеков, ежегодно проводимого в МГУ, и проекта измерения качества 3D-видео.

Anastasia Vsevolodovna ANTSIFEROVA – received her master degree in computer science from Moscow State University in 2018. She is a postgraduate student at Moscow State University and a member of Video Group in MSU Graphics&Media Lab. She is also a researcher at the Research Centre for Trusted AI of ISP RAS. Her research interests involve video codecs analysis and optimization, stereoscopic video subjective quality assessment. She is one of the contributors to MSU Video Codec Comparison Project and to the 3D video quality measurement project.

Дмитрий Сергеевич ВАТОЛИН закончил ВМК МГУ в 1996 году, кандидат физико-математических наук, заведующий лабораторией компьютерной графики ВМК МГУ, исследователь Центра доверенного искусственного интеллекта ИСП РАН. Читает курсы по компьютерной графике и методам сжатия и обработки видео с 1997 года. Создатель популярных сайтов, посвященных алгоритмам

обработки и сжатия видео. Специализируется на исследованиях в области алгоритмов сжатия видео, современных методах измерения качества и обработке цифрового видео. Руководил проектами с компаниями Intel, Cisco, Real Networks, Samsung, Huawei, Broadcom и некоторыми другими. С 2008 года занимается измерением и исправлением артефактов стерео, в том числе организовал проект измерения качества стереофильмов.

Dmitriy Sergeevich VATOLIN – Cand. Sci. (Phys.-Math.) from Moscow State University, head of the MSU Graphics & Media Lab and MSU AI Institute Video Analysis Lab, and a researcher at the Research Centre for Trusted AI of ISP RAS. His research interests include compression methods, video processing, 3D video techniques (depth from motion, focus and other cues, video matting, background restoration, high-quality stereo generation), as well as video quality assessment and robustness of modern video quality metrics. He is a key cofounder of the 3D video quality measurement project, his most known project is the annual MSU Video Codecs Comparison, that includes up to 25 modern codecs compared subjectively and objectively in several nominations with detailed 20000+ charts.