ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2022, № 2, с. 86–99

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.932.2, 004.852

ДЕТЕКТИРОВАНИЕ ПОДДЕЛОК В МОБИЛЬНЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ ПО ЛИЦУ ПРИ ПОМОЩИ СТЕРЕОКАМЕРЫ[1]

© 2022 г. Ю. С. Ефимов[a],*, И. А. Матвеев[b],**

[a] *МФТИ, Долгопрудный, МО, Россия*

[b] *ФИЦ ИУ РАН, Москва, Россия*

**e-mail: yuri.efimov@phystech.edu*

***e-mail: matveev@ccas.ru*

Поступила в редакцию 14.10.2021 г.
После доработки 16.10.2021 г.
Принята к публикации 29.11.2021 г.

Предложен метод обнаружения спуфинг-атак в системах распознавания по лицу на мобильных устройствах. Метод способен работать в реальном времени на устройствах с ограниченными вычислительными возможностями, различными условиями регистрации и с использованием штатной стереокамеры, особенностью которой является малый стереобазис. Метод основан на сверточной нейронной сети и многоцелевом обучении, предложена специальная функция потерь. Рассмотрены следующие типы подделок: высококачественные распечатанные изображения лиц, цифровые фотографии и видеопоследовательности, показываемые на экранах высокого разрешения. Полученный алгоритм протестирован на нескольких наборах стереоизображений как общедоступных, так и собранных вручную, отличающихся большим разнообразием условий съемки.

DOI: 10.31857/S0002338822020068

Введение. Технология распознавания по лицу получила распространение в разных областях, от крупномасштабных систем видеонаблюдения до мобильных устройств. Важным преимуществом этой технологии является возможность применения практически без взаимодействия с субъектом [1]. Как и другие биометрические модальности, распознавание лиц уязвимо по отношению к предоставлению подделок или *спуфингу*. В отличие от систем, использующих трудно воспроизводимые биометрические признаки, такие, как рисунок отпечатка пальца или текстура радужки, изображение лица человека несложно получить и подделать. Множество систем распознавания лиц использует изображения в видимом диапазоне, что позволяет осуществлять спуфинг-атаку на такую систему при помощи простой качественной фотографии, показываемой на цифровом экране или распечатанной на принтере высокого разрешения.

В настоящее время широко распространены социальные сети и сервисы по обмену фотографиями и видеозаписями, поэтому получить набор изображений лица практически любого человека не составляет труда. Помимо этого, изображения лица возможно получить и при помощи скрытой фото- или видеосъемки, которая сейчас доступна в высоком качестве каждому имеющему современный смартфон. Таким образом, уровень безопасности систем распознавания по лицу в первую очередь определяется устойчивостью к предъявлению подделок, следовательно, задача определения живости или *антиспуфинга* является актуальной.

1. Современное состояние области исследования. Подделки лица человека можно разделить на три группы по способу их создания: распечатки изображения лица, цифровые изображения или видео и лицевые маски [2]. Сложность детектирования подделок во многом определяется качеством используемых материалов и устройств. Важнейшим принципом обнаружения подделки лица является определение трехмерных характеристик видимой сцены. Первые два способа здесь отсекаются, третий становится значительно более трудоемким.

Описанные в литературе методы антиспуфинга можно разделить на две группы: использующие дополнительное оборудование (сенсоры глубины [3], камеры в ближнем инфракрасном

[1] Работа выполнена при финансовой поддержке РФФИ (гранты № 19-07-01231, 19-31-90167).

диапазоне (ИК-камеры) [4], тепловые камеры [5]) и основанные исключительно на программной обработке входного изображения. Методы первой группы позволяют решать задачу детектирования подделок с высокой точностью, однако их применение на практике существенно увеличивает стоимость системы. Вторую группу можно разделить на две подгруппы: кооперативные и некооперативные методы. Кооперативные методы требуют выполнения определенных движений лицом и/или его частями в соответствии с запросом системы, что повышает уровень безопасности, но раздражает пользователя и увеличивает время отклика системы.

Системы биометрической идентификации в мобильных устройствах, таких, как смартфоны и ноутбуки, должны иметь малое время отклика, возможность работать на ограниченных вычислительных ресурсах, допускать применение в разнообразных и неконтролируемых условиях съемки. В случае с изображением лица человека в видимом спектре изменчивость условий съемки делает возможным появление практически идентичных низкокачественных изображений настоящих лиц и подделок [2]. Эти факторы существенно ограничивают набор подходов к решению задачи.

Большинство представленных на рынке мобильных устройств с системой распознавания по лицу, дающей доступ к личной информации пользователя и осуществлению платежных операций, оборудовано дополнительными датчиками [6], [7] для обеспечения высокого уровня безопасности против спуфинг-атак, как правило сенсорами определения глубины. При этом лишь сравнительно небольшая группа устройств [8], [9] снабжена парой фронтальных камер, позволяющих оценивать глубину снимаемых сцен алгоритмами стереозрения. По сравнению с большинством более совершенных датчиков, используемых для оценки глубины сцены, дополнительная фронтальная камера вносит небольшую добавочную стоимость в систему по ряду причин. Во-первых, упомянутые сенсоры применяют технологию активной подсветки для получения карты глубины, что требует установки источника этой подсветки и приемника – дополнительной камеры, зачастую восприимчивой к ИК-излучению, что делает ее совмещение с основной фронтальной камерой невозможным. Во-вторых, два добавочных датчика требуют для установки дополнительное пространство на передней панели мобильного устройства, большую часть которой занимает сенсорный дисплей. Поэтому разумно исследовать возможности применения фронтальных стереокамер мобильных устройств для решения задачи антиспуфинга в системах распознавания по лицу.

Задача построения карты глубины сцены по изображениям стереокамеры является классической [10]. Ее решение, как правило, состоит из нескольких этапов: калибровка стереопары, построение карты смещений (диспаратности) и последующее построение карты глубин с учетом параметров калибровки.

Более современный подход – применение сверточных нейронных сетей [11–13]. При наличии достаточно большого по объему и разнообразию набора входных изображений можно решать как задачу оценки смещений [11], так и извлечения глубины изображения в разнообразных условиях освещенности [12]. Недостатками этих подходов применительно к их использованию в мобильных устройствах являются высокие вычислительные затраты и необходимость большого стереобазиса, такого, как, например, в беспилотных автомобилях. Описаны менее ресурсоемкие подходы [14]. Для предсказания глубины сцены авторы предлагают задействовать как информацию от мобильной стереопары с малым расстоянием между центрами камер, так и от подсистемы фазового фокуса одной из камер (PDAF, phase detection auto focus).

Большинство описанных в литературе методов, использующих стереоизображения для антиспуфинга, так или иначе пытаются извлечь информацию о глубине изображения или видео лица человека для определения его живости [15]. Эти подходы имеют преимущество перед однокадровыми, особенно для изображений, полученных в разнообразных условиях съемки. Как правило, такие алгоритмы основаны на классических или нейросетевых классификаторах, обрабатывающих признаки карты смещений или карты глубин. В [16] авторы предлагают способ построения приблизительной карты диспаратности при помощи небольшой сверточной нейронной сети, предобученной на целевом домене. После этого этапа обучается вторая нейронная сеть, использующая признаковые описания, которые получены от первой.

В [17] для предсказания метки класса предлагается нейронная сеть, построенная по принципу автокодировщика и состоящая из двух частей. Кодирующая часть нейронной сети извлекает промежуточные признаки, которые затем используются декодирующей частью для регрессии значений диспаратности. Результат обработки входных изображений декодером затем подается в небольшую сверточную нейронную сеть для классификации. Обе части нейронной сети обучаются совместно как на регрессию истинных значений диспаратности, так и на предсказание пра-

вильной метки класса, что повышает обобщающую способность полученной нейронной сети. Аналогичный подход был применен в ряде работ по детектированию подделок и привел к повышению точности итоговых решений [18]. Недостатком этого подхода является вычислительная сложность, поскольку для предсказания метки класса требуется пропустить входную пару изображений через обе части нейронной сети.

Все упомянутые работы используют обучающие выборки, полученные при помощи стереокамер с большим стереобазисом (более 4 см), что повышает устойчивость и точность восстановления трехмерных признаков. Однако типичные стереокамеры мобильных устройств имеют расстояния между центрами сенсоров не более 2 см.

В работе предлагается алгоритм определения живости лица на стереоизображении, основанный на применении сверточной нейронной сети. Используется нейронная сеть с небольшой вычислительной сложностью, обученная на парах изображений, полученных штатной стереокамерой мобильного устройства. Предлагаемый метод протестирован на наборе изображений, зарегистрированных при большом разнообразии условий освещенности.

2. Описание предлагаемого метода. Решение задачи основано на сверточной нейронной сети, входом которой является стереопара изображений. Используется информация о положении лица на одном из изображений.

2.1. Постановка задачи. Входными данными для предлагаемого метода является пара изображений, про которые известно, что на них лицо содержится целиком. Определение живости осуществляется после этапа образмеривания, т.е. когда найдены область лица и координаты глаз. Без ограничения общности будем предполагать, что пара изображений I_0 и I_1 есть квадратные растры размером $W \times W$ пикселей, индекс 0 соответствует левой камере, 1 – правой. Обозначим координаты центров правого и левого глаз на изображении (x_R, y_R) и (x_L, y_L) соответственно. Каждая пара изображений имеет также бинарную метку класса: $y = \{0;1\}$, где “1” обозначает “настоящее лицо”, “0” – “подделка”.

Решается задача классификации объектов, т.е. строится отображение:

$$a : (I_0, I_1; w) \mapsto y, \tag{2.1}$$

где w – набор параметров алгоритма. Для оценки точности бинарной классификации существуют известные характеристики: TP (true positive) – доля объектов класса “1”, которым была корректно присвоена метка “1”; FN (false negative) – доля объектов класса “0”, которым была некорректно присвоена метка “1”; FP (false positive) – доля объектов класса “1”, которым была некорректно присвоена метка “0”; TN (true negative) – доля объектов класса “0”, которым была корректно присвоена метка “0”.

Используются следующие меры качества, предписанные в стандарте [19]:

CCR (correct classification rate) – доля корректно классифицированных объектов:

$$\mathrm{CCR} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}; \tag{2.2}$$

APCER (attack presentation classification error rate) – доля изображений подделок, классифицированных как содержащие живое лицо:

$$\mathrm{APCER} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}; \tag{2.3}$$

BPCER (bona-fide presentation classification error rate) – доля изображений настоящих лиц, классифицированных как содержащие подделку:

$$\mathrm{BPCER} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}; \tag{2.4}$$

EER (equal error rate) – уровень ошибки классификатора при некотором пороговом значении решающего правила, при котором значения ошибок обоих видов равны.

2.2. Предварительная обработка изображений. Большое значение для качества работы классификатора имеет предварительная подготовка данных, минимизирующая различия в условиях регистрации. Здесь рассматриваются ректификация, определение ориентации лица и выбор цветового пространства.

Калибровка и ректификация. Первым этапом предобработки изображений стереопары является *ректификация* – приведение изображений к некоторому стандартному виду путем

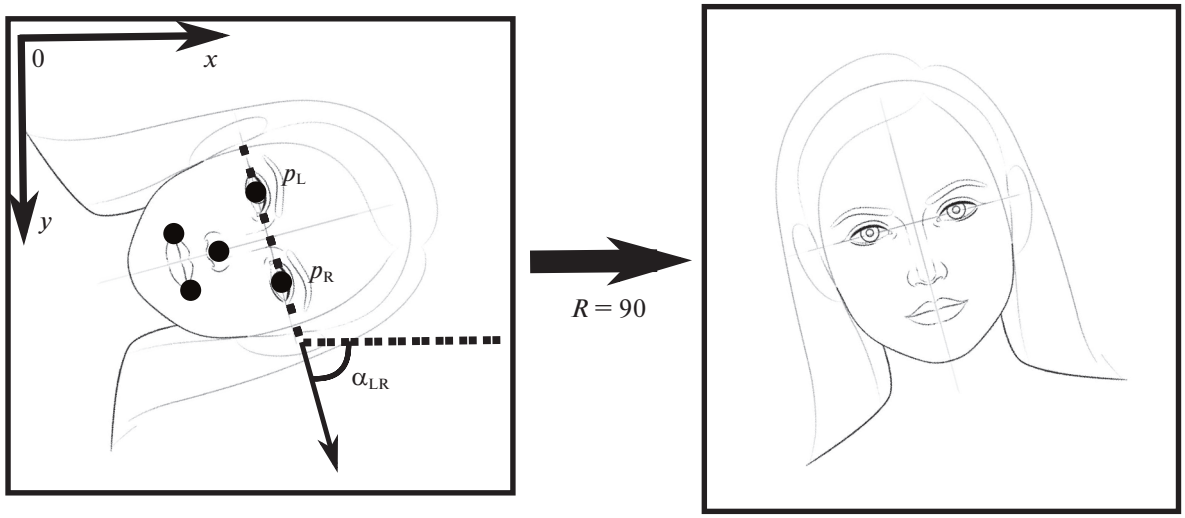


Рис. 1. Определение ориентации лица на изображении

компенсации искажений, вносимых инивидуальными особенностями камер. Ректификация производится на основании калибровочных данных, (матрица внутренних параметров камеры и коэффициенты дисторсии) могут быть извлечены из памяти устройства [20]. Каждая камера современных мобильных устройств калибруется на производственных линиях на специальной установке, как правило, еще до окончательной сборки корпуса. К сожалению, при сборке, транспортировке и эксплуатации положения сенсоров и/или линз камер может измениться, что приводит к несоответствию калибровки действительным параметрам камер [21]. Дефекты калибровки незаметны в большинстве приложений, однако являются существенными для фотограмметрии. Повторная калибровка на стороне пользователя нежелательна даже в автоматизированном режиме, поскольку это снижает удобство и вносит значительный риск некорректного выполнения.

По этой причине в данной работе используются изображения, полученные без ректификации.

Определение ориентации лица. Особенностью использования мобильных устройств является то, что их ориентация при распознавании может быть различной: портретной и ландшафтной. Детектирование подделок происходит после этапа образмеривания.

Среди точек, полученных при образмеривании, содержатся положения центров глаз $\vec{p}_R = (x_R, y_R)$ и $\vec{p}_L = (x_L, y_L)$, эти координаты можно использовать для определения ориентации входного растра. Угол наклона прямой, соединяющей зрачки к оси OX, равен

$$\alpha_{LR} = \left(\frac{y_R - y_L}{x_R - x_L} \right)$$

и определяет ориентацию R входного кадра (рис. 1):

$$R = \begin{cases} 0, & |\alpha_{LR}| < \dfrac{\pi}{4}, \\ 90, & \dfrac{\pi}{4} < \alpha_{LR} < \dfrac{3\pi}{4}, \\ 180, & |\alpha_{LR}| > \dfrac{3\pi}{4}, \\ 270, & -\dfrac{3\pi}{4} < \alpha_{LR} < -\dfrac{\pi}{4}. \end{cases} \tag{2.5}$$

Значение R задает угол, на который требуется повернуть исходный растр против часовой стрелки, чтобы ориентация лица на нем стала естественной, строго “подбородком вниз”.

Выбор цветового пространства. Цветовые каналы RGB-представления изображений сильно скореллированы, поэтому для повышения обобщающей способности нейронных сетей предлагается применять либо иные цветовые пространства, либо избавляться от цветности вовсе [22].

В работе изображения перед подачей в нейронную сеть преобразовываются к одноканальному (монохромному) представлению. К полученным парам растров применяется алгоритм блочного приведения гистограмм [23], чтобы избавиться от искажений, вносимых расхождением автоэкспозиций стереокамер.

2.3. Выбор функции потерь. Детектирование подделок – это задача бинарная классификация, при решении которой при помощи методов машинного обучения часто применяют логистическую функцию потерь или перекрестную кросс-энтропию:

$$L_0 = \frac{1}{K}\sum_{i=1}^{N}(-y^i \log a^i - (1-y^i)\log(1-a^i)) \to \min, \tag{2.6}$$

где $y^i \in \{0;1\}$ – метка класса, a^i – ответ алгоритма на i-м примере обучающей выборки, имеющей размер K.

Классификатором (2.1) является нейросеть $\mathbb{N}(I_0, I_1; w)$ с набором весов w. Ее предсказания соответствуют вероятности принадлежности лица на стереоизображении к положительному классу (в данном случае – к классу “настоящее лицо”):

$$a^i = \mathbb{N}(I_0^{(i)}, I_1^{(i)}; w) = P(y^i = 1; w). \tag{2.7}$$

Представим упомянутую сеть в виде композиции подсетей, осуществляющих извлечение признаков из пары изображений $\mathbb{N}_\text{f}(I_0, I_1; w_\text{f})$ и предсказание метки класса $\mathbb{N}_\text{o}(I; w_\text{o})$:

$$\mathbb{N}(I_0, I_1; w) = \mathbb{N}_\text{o}(\mathbb{N}_\text{f}(I_0^{(i)}, I_1^{(i)}; w_\text{f}); w_\text{o}). \tag{2.8}$$

В работе подразумевается, что при генерации метки класса итоговая модель должна опираться на отличия глубины сцен, содержащих настоящие и поддельные лица. При этом процедура минимизации функции потерь не гарантирует того, что модель научится извлекать релевантные и устойчивые признаки для корректных предсказаний на отложенных данных. Более того, как показывают эксперименты, оптимизация лишь перекрестной кросс-энтропии не обеспечивает хорошей обобщающей способности результата обучения в случаях ограниченных по разнообразию и размеру обучающих выборок.

В литературе описаны способы повышения обобщающей способности нейронных сетей за счет многоцелевого обучения [24]. Оптимизация осуществляется для суммы нескольких функций потерь, соответствующих разным, но связанным между собой подзадачам. В результате при прохождении входного сигнала через обученную нейронную сеть в ней возникают промежуточные представления, порождающие более устойчивое признаковое описание для решения каждой из подзадач, в том числе и основной. Обобщающая способность повышается за счет того, что процедура обучения не позволяет весам нейронной сети оказаться в тривиальном для данного набора данных локальном оптимуме.

Применение данного подхода описано для задачи детектирования подделок среди цветных изображений лиц [18, 25]. Полученные модели показывали лучшее качество по сравнению с их аналогами с единственной классификационной функцией потерь.

В работе используется вспомогательная функция потерь, которая позволяет нейронной сети извлекать информацию о глубине представленной на стереоизображении сцены. Предлагается добавить в модель подсеть $\mathbb{N}_\text{a}(I; w_\text{a})$, которая предсказывает карту принадлежности пикселей A к переднему плану с помощью признакового описания, полученного подсетью $N_\text{f}(.,.; w_\text{f})$. Каждый элемент A может принимать только значения 0 или 1. Примеры таких карт для истинного и поддельного лица показаны на рис. 2.

Для каждого пикселя первого из входных изображений I_0 с координатами $m \in [1; M]$, $n \in [1; N]$ нейронная сеть $\mathbb{N}_\text{a}(.; w)$ должна также предсказать вероятность:

$$b_{mn} = \mathbb{N}_\text{a}(\mathbb{N}_\text{f}(I_0, I_1; w_\text{f}); w_\text{a})_{mn} = P(A_{mn} = 1; w_\text{a}, w_\text{f}). \tag{2.9}$$

В таком случае внутренние представления обученной нейронной сети будут содержать в себе информацию, связанную с особенностями глубины изображений и величинами смещений между левым и правым изображениями стереопары. Для моделирования принадлежности пикселей

а *б*

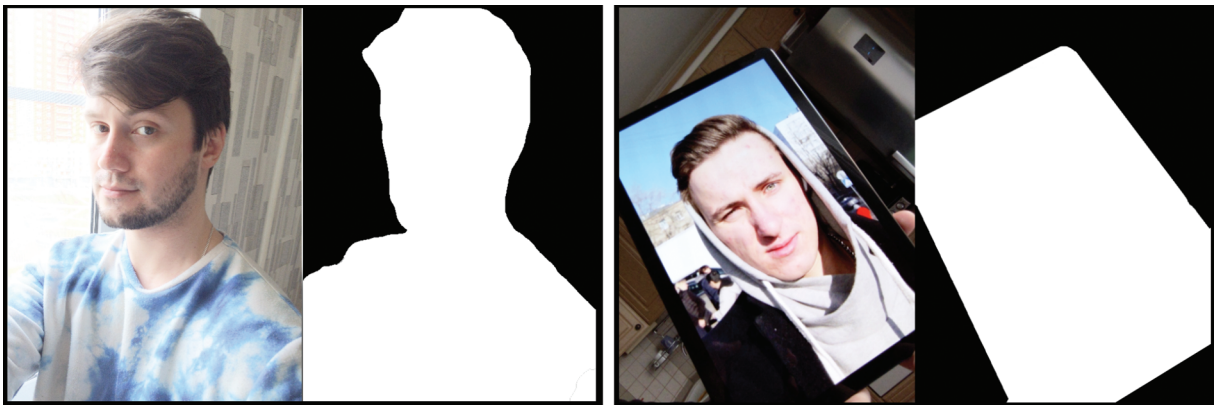


Рис. 2. Маски принадлежности пикселей к переднему плану: *а* – настоящее лицо; *б* – подделка

предлагается использовать сигмоидную функцию активации и для обучения применять логистическую функцию потерь (2.6) для каждого из пикселей по отдельности:

$$L_1^i = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(-A_{mn}^i \log b_{mn}^i - (1 - A_{mn}^i)\log(1 - b_{mn}^i)) \to \min, \tag{2.10}$$

$$L_1 = \frac{1}{K} L_1^i. \tag{2.11}$$

Итоговая функция потерь определяется как сумма: $L = L_0 + L_1$.

2.4. Архитектура модели. В качестве основы для построения нейронной сети была выбрана сравнительно легковесная архитектура семейства UNet [26] с добавлением остаточных блоков [27] для улучшения сходимости. Модель устроена по принципу автокодировщика, составленного из комбинаций сверточных блоков и операций сокращения пространственной размерности (пулинга). Общее строение представлено на рис. 3, архитектуры блоков кодирующей и декодирующей частей сети – на рис. 4. Каждый элемент блок-схемы содержит название операции, ее параметры и размер результирующего тензора. Символом *s*' обозначена величина смещения фильтра операции свертки. Большая часть сверточных блоков кодировочной части построена по принципу, описанному в [28] для повышения производительности архитектуры. Строение используемых сверточных блоков дано в табл. 1.

Пары изображений подаются в модель в исходной их ориентации (как они получены с сенсоров камер), чтобы сохранить горизонтальным направление смещений соответственных пикселей и упростить задачу сегментации переднего плана для декодирующей части. При этом промежуточное признаковое описание лица после обработки входного сигнала кодировщиком может быть представлено в некорректной ориентации (2.5). Чтобы упростить классификацию, имеет смысл повернуть это признаковое описание пространственно на угол, кратный $\pi/2$, таким образом, чтобы линия уровня глаз соответствовала ориентации $R = 0$. На рис. 3 эта операция обозначена как слой компенсации поворота.

2.5. Преобразование карты признаков. В сетях, построенных по принципу автокодировщиков, нейроны внутреннего представления обычно имеют большие рецептивные поля, охватывающие значительные связные области пикселей входного изображения. Такие представления с малой пространственной размерностью и большим количеством каналов содержат богатое агрегированное признаковое описание. В случае предсказания карты глубины или аналогичных задач с попиксельным предсказанием некоторых значений важно при построении ответа с помощью декодирующей части сети использовать внутреннее представление целиком. Однако задача детектирования подделок является локальной: содержательная часть признаков “живости” пространственно локализована в области лица на исходном растре с поправкой на операции пулинга. Отличия подделок от настоящих лиц содержатся именно в особенностях карты глубины вокруг лицевой области: у настоящего лица в этой области присутствует резкий перепад по отношению к заднему плану и плавные перепады на переднем плане, а у поддельного

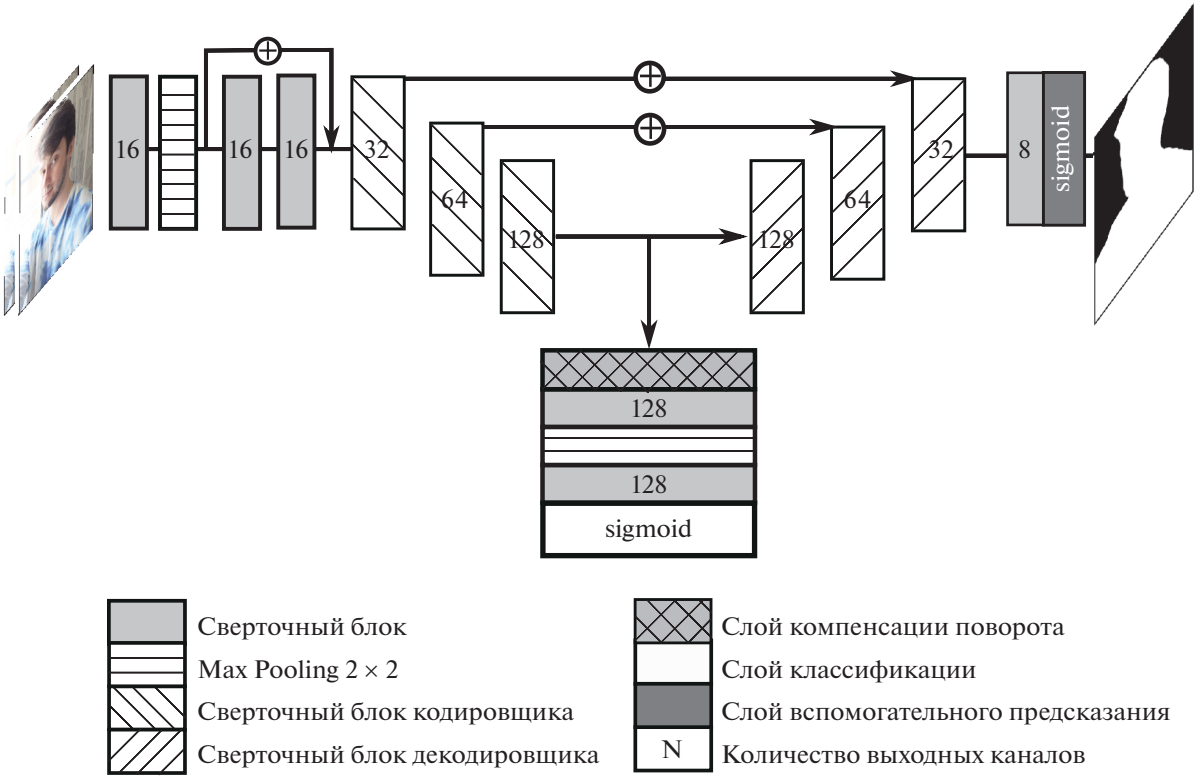


Рис. 3. Архитектура используемой нейронной сети

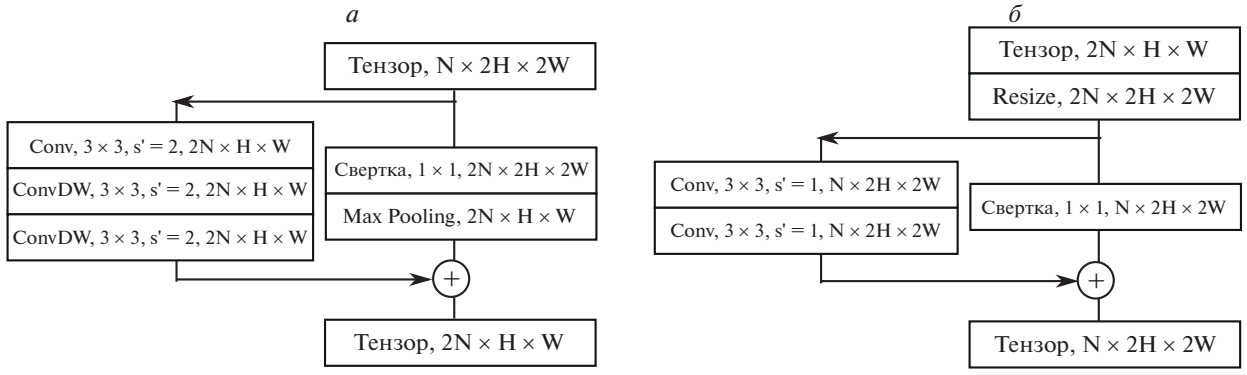


Рис. 4. Строение блоков: *а* – кодирующая часть; *б* – декодирующая часть

перепад к фону и на переднем плане отсутствует. Более того, учет информации от заднего плана может привести к переобучению в связи с ограниченным размером обучающей выборки.

При этом подача растра лицевой области напрямую в нейронную сеть нецелесообразна, так как часть информации о соотношении глубины фона и переднего плана может потеряться. Более того, лицо на изображении может занимать различную площадь ввиду разнообразия расстояний съемки, поэтому будет необходимо приведение входных данных к общему размеру, что может исказить исходную карту смещений между правым и левым растром в стереопаре. В работе предлагается отобразить область лица в промежуточном представлении входного сигнала нейронной сети в результирующий тензор T_o фиксированного размера $S \times S$ для его последующей обработки блоком предсказания метки класса. Для отображения используется билинейная интерполяция.

Таблица 1. Архитектура сверточных блоков

Слой	Размер ядра	Шаг	Размер входного тензора
Строение блока $Conv, s' = s$			
Свертка	3×3	s	$N \times K \times K$
Batch normalization	–	–	$N \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$
Активация ReLu	–	–	$N \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$
Строение блока $ConvDW, s' = s$			
Depth-wise свертка	3×3	s	$N \times K \times K$
Batch normalization	–	–	$N \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$
Активация ReLu	–	–	$N \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$
Свертка	1×1	1	$N \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$
Batch normalization	–	–	$M \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$
Активация ReLu	–	–	$M \times \left(\frac{K-3}{s}+1\right) \times \left(\frac{K-3}{s}+1\right)$

Область лица задается в данном случае как прямоугольник $C = (x, y, w, h)$, где x, y – положение центра прямоугольной области, w, h – ее ширина и высота соответственно. Значения параметров вычисляются из положения ключевых точек на лице:

$$x = \frac{1}{4}(x_L + x_R + x_{ML} + x_{MR}), \tag{2.12}$$

$$y = \frac{1}{4}(y_L + y_R + y_{ML} + y_{MR}), \tag{2.13}$$

$$w = h = 2\sqrt{(x_R - x_L)^2 + (y_R - y_L)^2}. \tag{2.14}$$

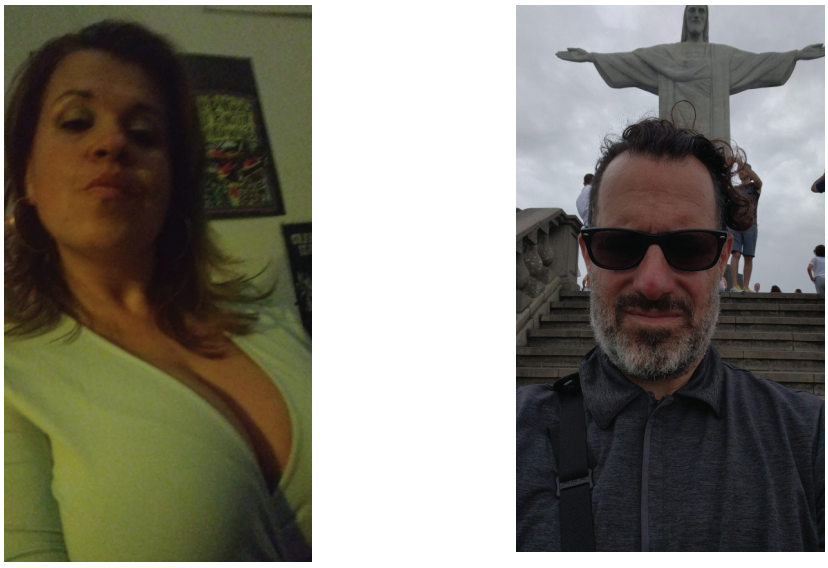
В данной работе разрешение результирующего представления области лица $S = 10$ выбрано с учетом средних параметров области интереса, определенных на обучающей выборке.

3. Численные эксперименты. Предложенный метод протестирован на различных наборах стереоизображений как общедоступных, так и собранных вручную. Большое внимание уделено получению изображений с разнообразными условиями регистрации.

3.1. Формирование базы изображений. В литературе описано несколько баз стереоизображений для задачи детектирования подделок, полученных при помощи полноразмерных стереопар с большим стереобазисом [16, 17]. Набора данных для мобильных приложений и поставленной задачи в открытом доступе авторы не нашли. Тем не менее, известна обширная база изображений Holopix [29], полученная при помощи мобильной стереокамеры и содержащая большое разнообразие типов сцен, среди которых присутствует класс “селфи”, соответствующий классу “настоящих лиц” в контексте данной работы. При съемке были использованы два вида сенсоров: со стереобазисом в 12 и 5 мм. Выборка содержит изображения различных разрешений, снятые в разнообразных условиях. Среди 50 тыс. растров этой выборки лишь 1052 можно отнести к классу “селфи”, который подразумевает наличие единственного лица в кадре и его расположение на расстоянии от 20 до 60 см до камеры (рис. 5). Для полуавтоматического отбора изображений применялся алгоритм детектирования лиц [30]. Упомянутый набор данных допустимо использовать в качестве отложенной тестовой выборки для проверки обобщающей способности алгоритма.

По причине нехватки открытых баз изображений был осуществлен дополнительный сбор данных при помощи Google Pixel 3 – одного из смартфонов с двойной фронтальной камерой,

а *б*



в

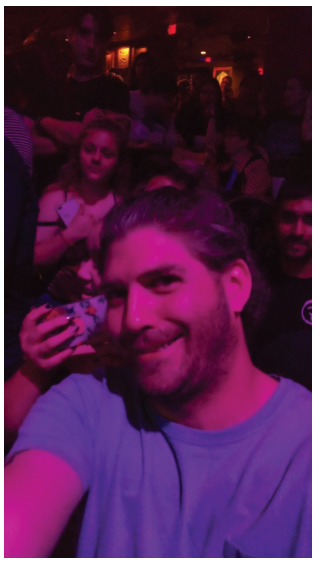


Рис. 5. Примеры изображений выборки Holopix50k

позволяющей получать изображения с обоих сенсоров одновременно. Выборка изображений лиц получена от 90 участников обоих полов. Каждому участнику предлагалось принять участие в съемке трех сценариев освещенности: естественное освещение в помещении (E1), яркая засветка с одной из сторон или всей сцены целиком (E2, E3 или E5) и съемка в полутемном помещении (E4). Примеры изображений даны на рис. 6.

Полученные фотографии частично были использованы для создания изображений подделок следующих типов: распечатка лица (PR, printed), лицо на экране высокого разрешения (SI, screen image) и лицо на небольшом дисплее мобильного устройства (SM, smartphone). Примеры изображений подделок даны на рис. 7. Сбор изображений подделок происходил как минимум в двух условиях освещенности: при достатке (E1 или E5) и недостатке света (E4).

Съемки каждого участника вживую и подделок его лица осуществлялись в двух ориентациях мобильного устройства: портретной и ландшафтной. Применялись два расстояния до сенсора камеры: порядка 25–30 см, что соответствует комфортному положению смартфона относительно глаз, и порядка 45–60 см – положение смартфона в вытянутой руке. Подробное описание

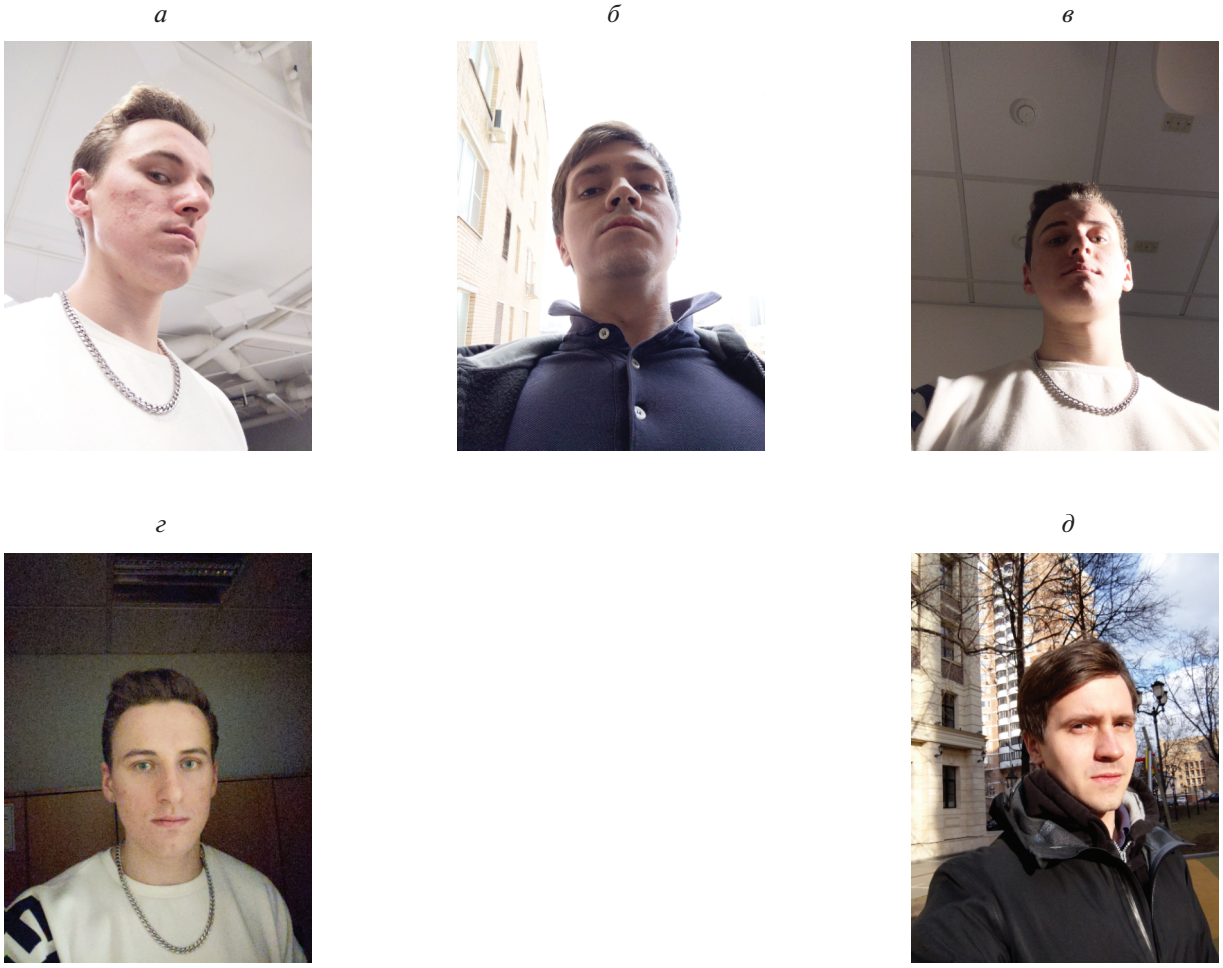


Рис. 6. Примеры условий освещенности: *а* – E1; *б* – E2; *в* – E3; *г* – E4; *д* – E5

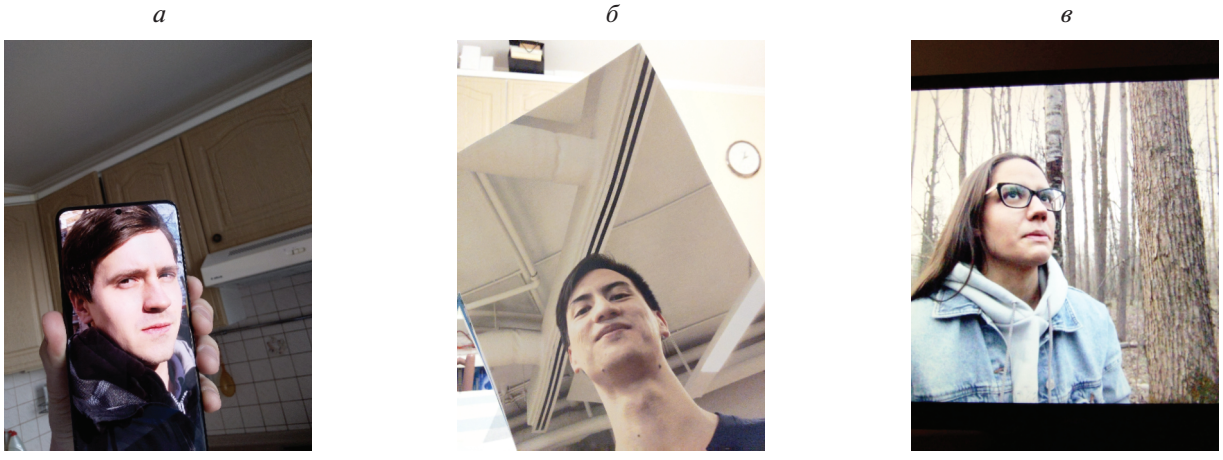


Рис. 7. Примеры изображений подделок: *а* – SM; *б* – PR; *в* – SI

полученного набора изображений приведено в табл. 2. Разбиение на обучающую и валидационную выборки осуществлялось в пропорции 7 к 3. При этом изображения одного и того же участника помещались лишь в одну из подвыборок.

Таблица 2. Описание использованной базы изображений

Тип освещенности	Количество изображений			
	Настоящие лица	Подделки		
		SM	PR	SI
E1	11326	5667	5844	7117
E2	7840	1839	8923	3006
E3	10980			
E5	3240			
E4	10335	4555	4617	4822
Всего	43721	12221	19384	15936

Положения лиц и ключевых точек на каждом из полученных изображений определены при помощи метода [30]. Для формирования бинарных карт принадлежности пикселей переднему плану, описанных ранее, использован один из методов вычисления оптического потока с его последующей бинаризацией [31]. Грубые ошибки определения оптического потока, возникающие вследствие некорректной работы камеры мобильного устройства и/или съемки в сложных условиях освещенности, исключены из обучающей выборки.

3.2. Границы применимости метода. Как известно, разрешающая способность стереопары по глубине dZ, согласно эпиполярной геометрии [32], зависит от нескольких параметров: величины стереобазиса B, фокусного расстояния используемых сенсоров f, погрешности измерения смещений d и самой глубины данной точки Z:

$$dZ = Z^2 \frac{d}{fB}. \tag{3.1}$$

Мобильное устройство Google Pixel 3 имеет следующие характеристики камер: $B = 10$ мм $= 0.01$ м, разрешение сенсоров $W \times H = 2448 \times 3264$ пикселя, $f = 4.5$ мм $= 0.0045$ м, апертура 1/1.8. Оценочный размер сенсора по разрешению и значению апертуры составляет порядка $\omega = 5$ мм $= 0.005$ м. Таким образом, один пиксель на выходном изображении имеет физический размер $\omega/W = 2 \times 10^{-6}$ м. При этом в целях повышения скорости обработки входных изображений нейронной сетью пространственное разрешение требуется сократить до некоторого

$$W_{\mathrm{CNN}} = \frac{W}{s}, \tag{3.2}$$

где $s > 1$ – коэффициент масштабирования, который можно определить в рамках поставленной задачи.

В данной работе предполагается, что обученная нейронная сеть должна уметь отличать настоящее лицо от плоского поддельного на расстоянии от 20 до 60 см при помощи информации о глубине сцены, которая содержится в стереоизображениях. Характерный размер головы человека можно определить как 20 см, поэтому величина dZ не должна превосходить это значение, чтобы потенциально извлекаемая карта глубины могла различать видимое лицо на фоне близкого плоского объекта позади, в худшем случае при $Z = 0.6$ м.

Точность определения смещений d задаётся равной 1 пикселю на растре пониженного разрешения W_{CNN}, т.е.

$$d = \frac{s\omega}{W} = 2 \times 10^{-6} s. \tag{3.3}$$

Требуется выполнение следующего неравенства:

$$Z^2 \frac{s\omega}{fBW} < dZ \leftrightarrow s < \frac{fBW \cdot dZ}{\omega Z^2}. \tag{3.4}$$

Подставляя ранее определенные значения, получаем

$$s \lessapprox 12.24. \tag{3.5}$$

В таком случае наименьший допустимый размер используемого изображения в пикселях равен

$$\frac{W}{s} = \frac{2448}{12.24} = 200.$$

Это значение было выбрано в качестве разрешения входных изображений для нейронной сети.

Ограничение сверху на разрешающую способность также определяет невозможность применения предлагаемого метода для детектирования подделок в виде плоских масок. Для детектирования этого типа спуфинга требуется разрешение $dZ \approx 5$ см $= 0.05$ м, дабы извлекать информацию о геометрии лица и его частей. Тогда требуется использовать коэффициент масштабирования $s_{\text{mask}} = \frac{1}{4}s \lessapprox 3.06$, что определяет минимальный размер изображения в 816 пикселей. Применение нейронных сетей в условиях видеопотока в реальном времени на маломощных мобильных вычислительных устройствах затруднительно.

3.3. Процедура обучения. Все эксперименты с обучением нейронных сетей проводились на обучающей части выборки с контролем на валидационной части. Модели обучались методом стохастического градиентного спуска с адаптивным моментом. Первоначальное значение темпа обучения составляло 0.001 и уменьшалось экспоненциально в 0.9 раза каждые 10 эпох. Обучение каждой модели проводилось на протяжении 128 эпох. Для предсказания метки класса при валидации использовались лишь кодирующий и классификационный блоки нейронной сети. Декодирующий блок не участвовал, поскольку его предназначение – лишь регуляризация обучения.

Во время обучения для повышения устойчивости модели к вариациям входных данных применялись аугментации случайной яркостной коррекции, наложения случайного пуассоновского шума, случайного аффинного поворота на угол до 20° относительно оптического центра изображения, случайного отображения по горизонтали и извлечения случайного региона фиксированного размера вокруг области лица, меньшего, чем размер исходного изображения.

Помимо этого, в отдельном эксперименте была применена операция случайного обнуления смещений между пикселями пары: для некоторых примеров вне зависимости от метки класса одно из них приравнивалось к другому, результату присваивалась метка “поддельное лицо” и вспомогательная маска (2.9) заполнялась нулями. Интуиция подобного подхода состоит в регуляризации процедуры обучения нейронной сети. В результате описанного преобразования построенные в модели признаковые описания должны опираться на особенности карт смещений, а не на текстурные характеристики растров. Нейронная сеть, обученная таким образом, помечена как “RandomZero”.

3.4. Сравнение модификаций предлагаемого подхода. Поскольку осуществить сравнение качества решения предлагаемого метода с аналогами затруднительно ввиду различий области применения и источников входных данных, решено осуществить сравнение с базовыми алгоритмами, не использующими стереоинформацию.

В работе рассматривается несколько способов построения нейросетевого классификатора для решения задачи антиспуфинга. Самый простой способ – использование лишь одного изображения из стереопары в черно-белом режиме для предсказания метки класса. Такая модель применена в качестве базового алгоритма с именем “Base”. Этот подход можно усложнить, добавив интерполяцию карты признаков вокруг региона лица. Модели с такой модификацией обозначены как “ROI”.

Основной способ предсказания метки класса – использование пары изображений в черно-белом режиме без добавления декодирующего блока (разд. 2.3) и без применения интерполяции признаков региона лица (разд. 2.5). Этот класс моделей имеет наименование “Stereo”. Далее этот подход можно развить, добавив соответствующие модификации, первую из которых предлагается обозначить через “Aux”.

Результаты вычислительных экспериментов с разными модификациями предлагаемого подхода даны в табл. 3. Основной мерой качества в задаче детектирования подделок считается значение равной ошибки классификации (EER). Значения APCER и BPCER отражают склонность моделей к присваиванию метки класса “1” или “0” на пороге принятия решения 0.5.

По итогам вычислительного эксперимента наилучшая точность решения на валидационной выборке достигается для модели типа “Stereo” со всеми упомянутыми выше модификациями. Базовые модели этого типа чаще присваивают входным изображениям метку класса “подделка”. При этом добавление модификаций, призванных получить более информативные признаки для

Таблица 3. Качество решения на валидационной выборке

Модель	Значения мер, %		
	APCER	BPCER	EER
Base	**0.12**	41.31	12.54
Base+ROI	0.58	15.06	4.82
Stereo	2.35	13.02	4.95
Stereo+Aux	0.89	**2.9**	1.89
Stereo+Aux+ROI	0.57	3.01	1.45
Stereo+Aux+ROI+RandomZero	0.23	5.24	**1.24**

решения поставленной задачи, действительно повышает обобщающую способность моделей, что отражается на итоговых мерах качества.

Модели типа “Stereo” позволяют получить лучшее решение задачи по сравнению с моделями типа “Base”. При этом добавление модификации “ROI” дает возможность повысить производительность до уровня базовой модели типа “Stereo” без модификаций. Модели “Stereo” используют больше информации при работе. Скорее всего, это связано с тем, что модификация “ROI” снижает склонность сети к переобучению на признаках заднего плана, а текстурной информации лицевой области растров достаточно, чтобы достичь сравнительно высокой точности решения на валидационной выборке.

В качестве отложенной тестовой выборки была использована подвыборка набора изображений [29]. Для оценки выбрана модель “Stereo + Aux + ROI + RandomZero”, порог принятия решения был принят равным порогу меры EER: 0.38. В результате 970 из 1052 пар растров было помечено алгоритмом как “настоящее лицо”. Это соответствует точности классификации в 92.2%. Среди ошибок классификации большую часть (51 пример) составляют пары, содержащие изображение лица на большом расстоянии и снятые на сенсоры со стереобазисом в 5 мм, что можно понять из визуализации карты смещений для этих примеров. Оставшиеся ошибочные предсказания содержат, напротив, лица, снятые с очень близкого расстояния (менее 20 см), и лица, снятые со значительными бликами от солнца в кадре.

3.5. Оценка производительности. Предлагаемая модель для классификации в режиме тестирования использует лишь кодирующий и предсказательный блок описанной нейронной сети. Суммарное количество операций умножения и сложения в данных блоках модели составляет порядка 63.2 MFlop. Медианное время выполнения на одном ядре процессора Snapdragon 888 составляет 65 мс. Применение 8-битной квантизации весов и активаций обученной нейронной сети методом [33] позволяет сократить время выполнения до 23 мс за счет применения целочисленной арифметики и оптимизации операций свертки.

Заключение. Предложен метод определения спуфинг-атак в мобильных системах распознавания по лицу с применением пары камер с малым стереобазисом. Он заключается в использовании сверточной нейронной сети небольшого размера, обученной со специальной функцией потерь. Предлагаемый подход достигает высоких показателей точности детектирования подделок, сравнимых с описанными в современной литературе аналогичными подходами, в том числе на данных открытой базы стереоизображений. От известных аналогов предлагаемый метод отличается малым временем выполнения на современных мобильных процессорах, поэтому может быть применен для детектирования подделок в биометрических системах с малыми вычислительными ресурсами.

СПИСОК ЛИТЕРАТУРЫ

1. *Galbally J., Marcel S., Fierrez J.* Biometric Antispoofing Methods: A Survey in Face Recognition // IEEE Access. 2014. V. 2. P. 1530–1552.
2. *Yu Z., Qin Y., Li X., Zhao C., Lei Z., Zhao G.* Deep Learning for Face Anti-Spoofing: A Survey // arXiv:2106.14948v1 [cs.CV] 28 Jun 2021.
3. *Sun X., Huang L., Liu C.* Multimodal Face Spoofing Detection via RGBD Images // Proc. Intern. Conf. Pattern Recognition. Beijing, China, 2018. P. 2221–2226.
4. *Song L., Liu C.* Face Liveness Detection Based on Joint Analysis of RGB and Near-Infrared Image of Faces // Electronic Imaging. 2018. V. 6. P. 373-1–373-6.

5. *Seo J., Chung I.-J.* Face Liveness Detection Using Thermal Face-CNN with External Knowledge // Symmetry. 2019. V. 11. 3. P. 360.
6. Apple Face ID Technology // Электронный ресурс https://support.apple.com/en-us/HT208108, дата обращения 11.10.2021.
7. Google Face Unlock Technology // Электронный ресурс https://support.google.com/pixelphone/answer/9517039, дата обращения 11.10.2021.
8. Google Pixel 3 Device Description // Электронный ресурс https://en.wikipedia.org/wiki/Pixel_3, дата обращения 11.10.2021.
9. Samsung Galaxy S10+ Device Description // Электронный ресурс https://en.wikipedia.org/wiki/Samsung_Galaxy_S10, дата обращения 11.10.2021.
10. *Лобанов А.Н.* Фотограмметрия. М.: Недра, 1984.
11. *Chang J., Chen Y.* Pyramid Stereo Matching Network // Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA, 2018. P. 5410–5418.
12. *Khamis S., Fanello S., Rhemann C., Kowdle A., Valentin J., Izadi S.* StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction // Proc. 15th Europ. Conf. Computer Vision. Munich, Germany, 2018. P. 596–613.
13. *Zhang F., Prisacariu V., Yang R., Torr P.H.S.* GA-Net: Guided Aggregation Net for End-to-End Stereo Matching // Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019. P. 185–194.
14. *Zhang Y., Wadhwa N., Orts-Escolano S., Haene C., Fanello S., Garg R.* Du2Net: Learning Depth Estimation from Dual-Cameras and Dual-Pixels // European Conf. Computer Vision. Lecture Notes in Computer Science / Eds. A.Vedaldi, H.Bischof, T.Brox, JM.Frahm. V.12346. Springer, Cham, 2020.
15. *Sun X., Huang L., Liu C.* Dual Camera Based Feature for Face Spoofing Detection // Communications in Computer and Information Science. 2016. V. 662. P. 332–344.
16. *Rehman Y.A., Po L.M., Liu M.* SLNet: Stereo Face Liveness Detection via Dynamic Disparity Maps and Convolutional Neural Network // Expert Systems with Applications. 2020. V. 142. P. 113002.
17. *Li Z., Yuan J., Jia B., He Y., Xie L.* An Effective Face Anti-Spoofing Method via Stereo Matching // IEEE Signal Processing Letters. 2021. V. 28. P. 847–851.
18. *Liu Y., Jourabloo A., Liu X.* Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision // IEEE/CVF Conf. Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018. P. 389–398.
19. International Organization for Standardization ISO/IEC 30107-1:2016. Information Technology – Biometric Presentation Attack Detection. Pt 1. Framework, 2016.
20. Android Camera API // Электронный ресурс https://developers.google.com/ar/reference/java/com/google/ar/core/CameraIntrinsics, дата обращения 11.10.2021.
21. *SiVerdi S., Barron J.T., Shao X.* Geometric Calibration for Mobile, Stereo, Autofocus Cameras // Proc. IEEE Winter Conf. Applications of Computer Vision. Lake Placid, NY, USA, 2016. P. 1–8.
22. *Atoum Y., Liu Y., Jourabloo A., Liu X.* Face Anti-spoofing Using Patch and Depth-based CNNs // Proc. IEEE Intern. Joint Conf. Biometrics. Denver, Colorado, USA, 2017. P. 319–328.
23. *Прэтт У.* Цифровая обработка изображений. М.: Мир, 1982.
24. *Ruder S.* An Overview of Multi-Task Learning in Deep Neural Networks // arXiv:1706.05098v1 [cs.LG] 15 Jun 2017.
25. *George A., Marcel S.* Deep Pixel-wise Binary Supervision for Face Presentation Attack Detection // Proc. Intern. Conf. Biometrics. Crete, Greece, 2019. P. 1–8.
26. *Ronnenberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Proc. 18th Intern. Conf. Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015. P. 234–241.
27. *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // Proc. IEEE Conf. Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016. P. 770–778.
28. *Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H.* Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications // arXiv:1704.04861v1 [cs.CV] 17 Apr 2017.
29. *Hua Y., Kohli P., Uplavikar P., Ravi A., Gunaseelan S., Orozco J., Li E.* Holopix 50k: A Large-Scale In-the-wild Stereo Image Dataset // arXiv:2003.11172v1 [cs.CV] 25 Mar 2020.
30. *Zhang K., Zhang Z., Li Z., Qiao Y.* Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks // IEEE Signal Processing Letters. 2016. V. 13. № 10. P. 1499–1503.
31. *Liu C.* Beyond Pixels: Exploring New Representations and Applications for Motion Analysis // Doctoral Thesis. Massachusetts Institute of Technology. 2009.
32. *Шапиро Л., Стокман Дж.* Компьютерное зрение. М.: Бином. Лаборатория знаний. 2006. 752 с.
33. *Bhalgat Y., Lee J., Nagel M., Blankevoort T., Kwak N.* LSQ+: Improving Low-bit Quantization through Learnable Offsets and Better Initialization // arXiv:2004.09576v1 [cs.CV] 20 Apr 2020.