Московский Государственный Универститет им. М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики.
Кафедра информационной безопасности.

Садовников Алексей Андреевич

Исследование существующих атак на видеопоследовательности

Курсовая работа

Научный руководитель:
Ильюшин Евгений Альбинович

Москва, 2021

Содержание

1. Введение

Широкое распространение недорогих и портативных устройств видеозахвата, таких как цифровые камеры и мобильные телефоны, в сочетании с заметным ростом использования камер наблюдения, привело к внезапному увеличению объема цифровых аудиовизуальных данных, которые генерируется каждый день. А также технологические достижения различных инструментов для редактирования видео и изображений достигли такого уровня, что подделка цифрового видео или изображения может быть легко осуществима без ухудшения их качества и без каких-либо визуальных доказательств. В последние года особенно сильно возросло количество поддельных видео. Видео и изображения, доступные на различных платформах социальных сетей, таких как VK, YouTube, Facebook и др. играют жизненно важную роль в научном развитии и социально-экономическом восприятии. Помимо этого, видео используются в различных приложениях, таких как юридические доказательства, видеоуроки, реклама, видеонаблюдение. Злоупотребление или распространение неверной информации может производится через видео, т.е видео для широкого просмотра, например на YouTube или на каких-то телевизонных телеканалах может быть подвергнуто фальфикации.

Пример подделки видео

Рис. 1: На картинке есть дерево

Рис. 2: На картинке дерево было удалено

Была произведена атака удаления объекта в исходной последовательности видеокадров и на второй картинке уже нет дерева. Если видеть кусок данной видеопоследовательности впервые, то уличить визуально в подделке довольно сложно.

2. Цель работы

Исследовать существующие атаки на видеопоследовательности и некоторые методы их обнаружения.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Исследовать какие существуют виды атак на видеопоследовательности.
2. Проанализировать методы обнаружения атак на видеопоследовательности.

3. Атаки на видеопоследовательности

Непрерывная видеопоследовательность - $V(x, y, t)$ скалярная функция координат x, y и времени t. $B(x, y, t)$ - вектор модификации. $M(x, y, t)$ - подделанное видео, которая определяется как: $M(x, y, t) = B(x, y, t) + V(x, y, t)$. Когда содержание информации, создаваемой данной видеопоследовательностью, злонамеренно изменяется, это называется подделкой видеоданных. Видеопоследовательность можно рассматривать как набор последовательных кадров с временной зависимостью в трехмерной плоскости.

3.1. Классификация атак

Есть несколько возможных атак, которые могут быть применены для изменения содержимого видеоданных. Вмешательство в видео может либо изменить его содержимое, либо повлиять на временную зависимость между кадрами.

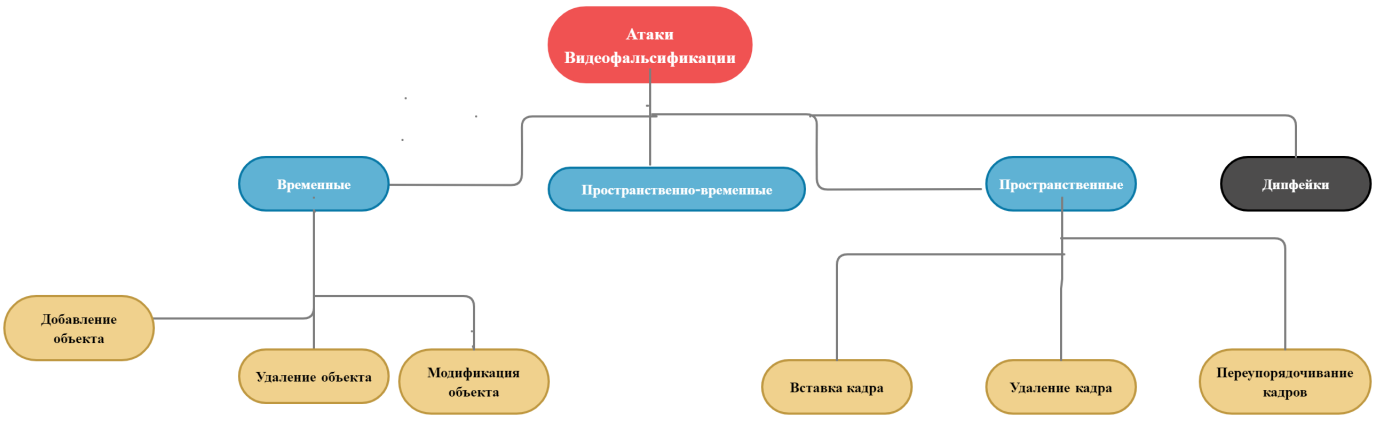

Рис. 3: Схема классификации

Атаки видеофальсификации разделяют на три основных области и еще одну отдельную дополнительную: 1) пространственная фальсификация 2) пространственно-временная фальсификация 3) временная фальсификация 4) дипфейки

3.2. Пространственное вмешательство

При пространственном редактировании вредоносные модификации изменяют содержимое одного или нескольких кадров.

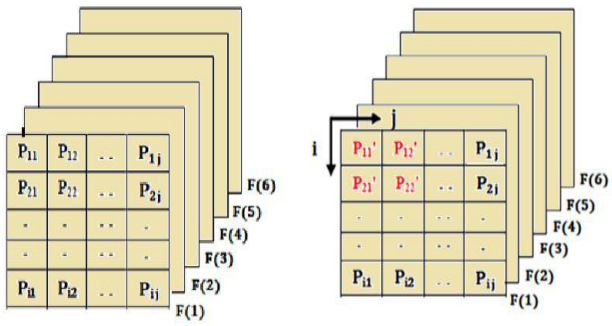

Рис. 4: Пространственное вмешательств

Внутрикадровое вмешательство показано на Рис. 4, на котором в кадре F(1) исходного входного видео V_0 пространственны искажены пиксели в левом верхнем углу матрицы для создания поддельного видео V_t. (i,j) обозначают высоту и ширину кадров входного видео соответственно.

Содержимое видеокадров обрабатывается как объекты. Объекты кадров можно разделить на два условных класса: объекты переднего плана и заднего плана.

Объекты переднего плана – это те, которые фиксируются как отдельные элементы в кадре без фона.

Фоновый объект – это фоновая часть часть кадра, за исключением всех объектов переднего плана.Атаки такого типа могут быть эффективно выполнены с помощью программного обеспечения для редактирования видео, например с помощью Adobe After Effects. К различным типам атак пространственного вмешательства относятся добавление объекта, удаление объекта и модификация объекта.

3.2.1. Атака с добавлением объекта

Эта атака может выполняться с объектами обоих типов, объектами заднего плана и объектами переднего плана. Копия-подделка с перемещением или подделка с копированием и вставкой является примером атаки с добавлением объектов. Используя эту атаку,

злоумышленник может вставить или удалить объект в сцене, изображенную на видеокадрах.

Рис. 5: Атака с добавлением объекта

На картинке показан пример подделки «копирование-перемещение», в котором дополнительное дерево в качестве объекта переднего плана копируется из исходного кадра и добавляется в другое место в том же кадре.

3.2.2. Атака с удалением объекта

В такой атаке удаляются объекты на кадрах видео. Эта атака может выполняться как с фоновым объектом, так и с объектом переднего плана. Для восстановления удаленных или поврежденных областей визуально правдоподобно можно использовать метод «рисование». Его можно использовать обычно двумя способами. Либо удаленные области заполняются с помощью образцов текстур, либо наиболее взаимосвязанные блоки из смежных по времени кадров используются для заполнения удаленной области. Еще есть подвид атаки - высококвалифицированная обрезка, при которой кадры видео обрезаются, чтобы удалить доказательства совершения подмены видео, а затем увеличивают поврежденные кадры, чтобы сохранить постоянное разрешение по всему видео.

Рис. 6: Атака с удалением объекта

На картинке показан пример атаки удаления объекта, при которой объект переднего плана

удаляется из исходного видеокадра.

3.2.3. Атака с модификацией объекта

При атаке модификации объекта существующий объект кадра может быть изменен таким образом, что первоначальная идентичность этого объекта теряется. Эта атака может быть выполнена как с фоновыми, так и с передними объектами. Возможны изменения размера, формы объекта, цвета объекта, а также с помощью дополнительных эффектов изменение характеристики объектов и его связь с другими объектами. Эти атаки выполняются на уровне пикселей. Эта атака считается довольно сложной для обнаружения. Фактически, DeepFakes - частный случай этой атаки.

Рис. 7: Атака с модификацией объекта

3.3. Временное вмешательство

Временная фальсификация видео – тип вмешательства, применяемый к видеокадрам, влияющий на временную последовательность визуального контента.Основное внимание уделяется временной зависимости. К распространенным атакам этого типа относят: вставку кадра, удаление кадра, переупорядочивание или перестановка кадров.

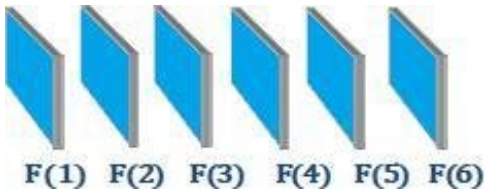

Рис. 8: Последовательность 6 кадров в исходном видео V_0

3.3.1. Атака с вставкой кадра

При этой атаке вставляются дополнительные кадры из другого видео, которое имеет те же статистические свойства,кадры намеренно вставляются в некоторые произвольные места в данном видео. При этом счетчик кадров увеличивается при вставлении новых кадров в исходное видео. Основная цель такой атаки – замаскировать исходный контент и предоставить ошибочную информацию.

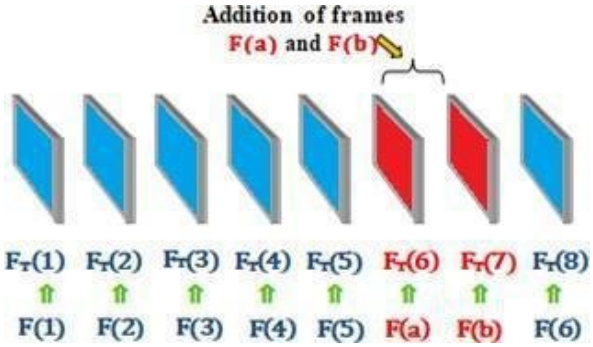

Рис. 9: Атака с вставкой кадра

На картинке пример, в котором два квадра F(a) и F(b) вставлены в случайном месте в исходное видео для создания искаженного видео, состоящего из восьми кадров.

3.3.2. Атака с удалением кадра

При атаке с удалением кадра, кадры удаляется намеренно из разных мест или могут быть удалены из определенного места. Количество кадров уменьшается, когда кадры удаляются из исходного видео. В зависимости от мотива это обычно выполняется при видеонаблюдении, когда злоумышленник хочет убрать свое присутствие на видео.

На картинке показан типичный пример атаки с удалением кадра из видеопоследовательности длиной 6 кадров, при котором кадры, помеченные F(3) и F(4) удаляются из исходного видео для создания поддельного видео, состоящего всего из четырех кадров.

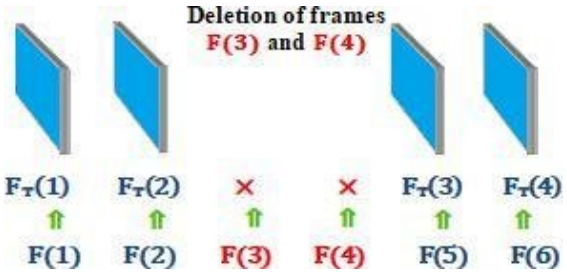

Рис. 10: Атака с удалением кадра

3.3.3. Атака с перестановкой кадров

При этой атаке кадры переупорядочиваются или перемешиваются таким образом, что фактическая последовательность видеокадров перемешивается и видео генерирует ошибочную информацию по сравнению с исходным видео. При этом счетчик кадров остается неизменным.

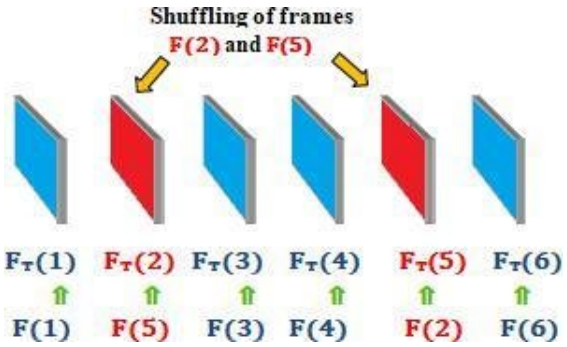

Рис. 11: Атака с перестановкой кадров

Типичный пример атаки с переупорядочиванием кадров, в которой 2 и 5 кадр меняются местами.

3.4. Пространственно-временная фальсификация видео

Пространственно-временная фальсификация видео - сочетание пространственного и временного вмешательства, можно найти как межкадровые, так и внутренние подделки. Система аутентификации должна быть достаточно надежной, чтобы распознать оба вида взлома.

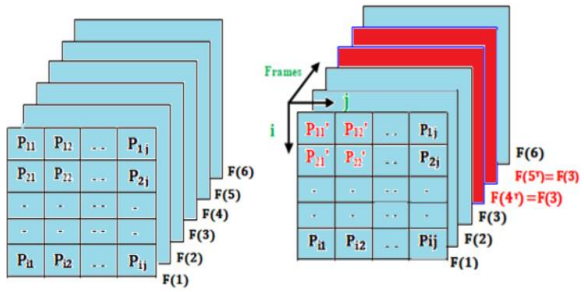

Рис. 12: Пространственно-временное вмешательство

i,j обозначают высоту и ширину входной последовательности видеокадров. Видно, что в результате временного вмешательства были переставлены кадры: а именно в результате временного вмешательства на место 4 и 5 кадров поставлен 3-ий кадр и в результате пространственного вмешательства в 1-ом кадре были изменены пиксели в левом верхнем углу матрицы пикселей.

3.5. Вмешательство в видеопоследовательности по уровням

Вмешательство в видео различают по уровням: на уровне пикселей, на уровне блока, на уровне кадра.

1) Подделка на уровне пикселей. Наименьший уровень, на котором может быть вмешательство в видеопоследовательность. Многие обычные операции обработки видео выполняются на уровне пикселей. Система аутентификации видео должна быть достаточно надежной, чтобы различать фальсификацию на уровне пикселей и нормальное обработку видео. На уровне пикселей обычно выполняется пространственное вмешательство

2) Подделка на уровне блоков. Указанная конкретная область в кадре видео называется блоком. Содержимое видеокадра рассматривается как множество блоков, над которыми выполняются вмешательства, блоки могут быть преобразованы, обрезаны, модифицированы или заменены. На уровне блоков обычно выполняется пространственное вмешательство.

3) Подделка на уровне кадра. Это вставка кадра, удаление кадра, перестановка кадров – распространенные атаки взлома видеопоследовательности. На уровне кадра обычно выполняется временное вмешательство.

4. Методы обнаружения атак на видеопоследовательности

Методы обнаружения подделки видео можно разделить на две основные категории: активные методы обнаружения подделки видео и методы пассивного обнаружения подделки видео.

4.1. Активные методы обнаружения

Активный метод обнаружения подделки видео использует предварительно встроенные скрытые данные, такие как цифровая подпись или водяной знак, для проверки подлинности и целостности цифровых видео.

4.1.1. Недостатки активного метода обнаружения

У активного подхода есть недостатки:

1. На этапе сбора данных требуется специальное оборудование, такое как оборудованные камеры для вставки водяных знаков или цифровой подписи в видео.

2. Многочисленные методы шифрования могут предотвратить несанкционированный доступ и манипулирование содержанием видео, тем не менее, эти методы шифрования не могут помешать владельцу цифрового видео

3. Такие факторы как сжатие, масштабирование, шум и т.д влияют в худшую сторону по надежности водяных знаков и цифровых подписей.

4.1.2. Цифровая подпись

В 1976 году Диффи и Хеллман ввели цифровую подпись для аутентификации мультимедийных данных и проверки их целостности. Цифровые подписи могут быть сохранены двумя разными способами для целей аутентификации. Либо его можно сохранить как независимый файл, либо в поле заголовка сжатой исходной информации. Это оказывается лучше, потому что цифровая подпись остается неизменной даже при изменении значений

пикселей изображений или видео, и это дает лучшие результаты. При такой аутентификации, цифровая подпись не может быть подделана, поскольку цифровая подпись подписывающего лица зависит от содержания данных и от некоторой секретной информации, которая известна только подписывающей стороне.

4.1.3. Интеллектуальные методы аутентификации

Интеллектуальные методы используют базу данных видеоклипов для аутентификации видео. База данных состоит как из подделанных, так и из подлинных видео. Основное преимущество интеллектуальной техники перед другими методами, такими как цифровая подпись и водяной знак, состоит в том, что она не требует какой-либо процедуры внедрения водяного знака или вычисления и хранения какого-либо секретного или открытого ключа.

Вычисляется информация о локальной относительной корреляции и классифицирует видео как поддельное или не поддельное видео. Используются алгоритмы машинного обучения - метод опорных векторов.

Алгоритм выполняется в два этапа:

1) Этап обучения

2) Этап обнаружения и классификации подделки

На этапе обучения алгоритм использует обучающие датасеты, помеченных вручную. Если видео в обучающих данных фальсифицировано, то назначается метка -1, а если подлинное, то +1. Из обучающих видео относительная корреляция информация между двумя соседними кадрами видео, затем вычисляется относительная корреляционная информация RC для всех смежных кадров видео с помощью: $RC = \frac{1}{m}\sum_{i=1}^{m} L_i$, где L_i - локальная корреляция между двумя кадрами, m - количество соответствующих угловых точек в двух кадрах. Информация о локальной корреляции RC вычисляется для каждого видео и RC с меткой информации обо всех обучающих видеоданных предоставляется в виде входных данных для метода опорных векторов. Со всей этой информацией обо всех видео с помощью метода опорных векторов алгоритм обучен определять является ли видео поддельным или нет. Концом обучения является верное построение гиперплоскости, отделяющая поддельные видео от подлинных.

4.1.4. Водяные знаки

Помимо проверки целостности цифровых данных и выявления злонамеренных манипуляций, водяные знаки могут использоваться для аутентификации производителя или автора контента. Водяные знаки могут быть встроены в видео, не изменяя фактического значения содержания данных. Полезная особенность водяных знаков заключается в том, что их можно встраивать без значительного ухудшения качества видео. Поскольку водяные знаки встроены в содержимое видеоданных, после изменения данных эти водяные знаки также будут изменены, это поможет использовать водяные знаки для проверки целостности мультимедийных данных. Техника нанесения водяных знаков на видео делится на два сегмента:

1. Встраивание или кодирование водяного знака во входное видео Процесс нанесения водяных знаков или кодирования видео выполняется на стороне источника. В этом процессе водяной знак встраивается во входное видео с использованием любого алгоритма водяных знаков. Весь процесс можно рассматривать как функцию $V_w = E(V_{in}, W, K)$, которая сопоставляет входное видео V_{in}, водяной знак W и ключ K для вывода видео с водяными знаками V_w

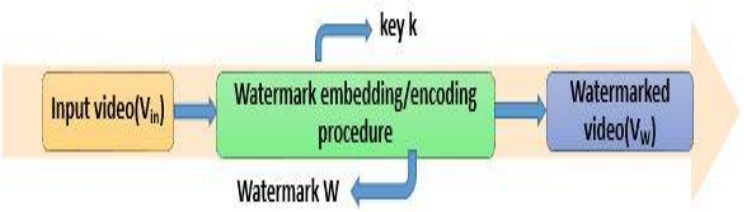

Рис. 13: Схема: встраивание водяного знака

2. Извлечение или декодирование водяного знака из видео. Процесс декодирования или извлечения водяных знаков из видео с водяными знаками является процессом, обратным алгоритму встраивания $V_{in} = D(V_W, W, K)$

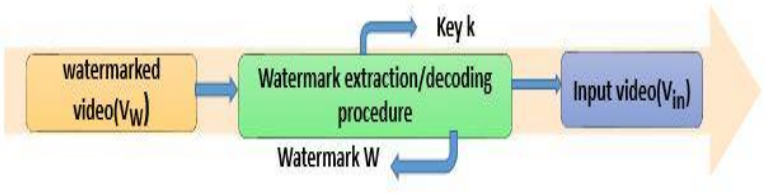

Рис. 14: Схема: извлечение водяного знака

4.1.5. Другие методы аутентификации

Помимо цифровой подписи, водяных знаков и интеллектуальных методов есть и иные способы аутентификации подделки видео. Одним из таких способов, является схема аутентификации для цифрового видео, основанная на траектории движения и совместном использовании криптографического секрета.В этой схеме видео сначала сегментируется на кадры, затем все кадры видеопоследовательности отображаются на траекторию в пространстве признаков, по которой вычисляются ключевые кадры видео. После того, как ключевые кадры вычислены, секретный кадр вычисляется на основе информации о ключевых кадрах видеопоследовательности. Эти секретные кадры используются для построения иерархической структуры, после чего получается окончательный главный ключ. Этот мастер-ключ используется для определения подлинности видео. Любые изменения в кадре или в важном содержании кадра будут отражены как изменения в вычисленном главном ключе. Траектория строится по гистограмме кадров. Как только ключевые кадры вычислены, они используются для вычисления секретного кадра путем экстраполяции. Интерполяционный полином $f(x)$ вычисляется с использованием ключа кадра. $\sum_{j=1}^{n+1}\prod_{i=1}^{n+1}\frac{x-x_i}{x_j-x_i}I_j$ - интерполяционная формула Лагранжа, где x_i - позиция относится к каждому ключевому кадру, а I_i - значение пикселя ключевых кадров. Используя это уравнение и экстраполяцию, а кадр при $x = 0$, считается секретным ключом. С помощью этой схемы любое видео может быть аутентифицировано путем сравнения его вычисленного мастер-ключа с исходным мастер-ключом. Это сравнение может быть выполнено с помощью общей меры косинусной корреляции, заданной $sim = \frac{I_0 I_N}{|I_0 I_N|}$, где I_0 - исходный главный ключ и I_N - новый главный ключ, рассматриваемые как векторы. Значения сходства будет находиться в диапазоне $[0, 1]$. Если $sim = 1$, то два главных ключа одинаковые, а если $sim = 0$, то два главных ключа полностью разные.

4.2. Пассивные методы обнаружения

Иное название пассивных методов - методы слепого обнаружения. Эти методы, которые можно использовать для проверки подлинности видео вне зависимости от предварительно встроенных или предварительно извлеченных данных. Обнаружение пассивного вмешательства включают обнаружение редактирования на основе камеры, на основе артефактов кодирования, на основе несоответствий в содержании, обнаружения копирования-перемещения в видео.

4.2.1. Преимущества пассивного подхода

1. Для обнаружения редактирования не требуется предварительно встроенной информации в видео, она зависит только от доступного измененного видео и его внутренних функций.

2. Для обнаружения редактирования видео не требуется специального оборудования.

4.2.2. Обнаружения редактирования пассивными методами

- Обнаружение редактирования видео на основе камеры.

  На записанных видеороликах видеокамеры оставляют характерный отпечаток пальца. Этот отпечаток пальца используется не только для идентификации устройства, но также может использоваться для обнаружения подделки видео. Прямое применение метода снятия отпечатков пальцев (неоднорость фотоответа) , в котором используются шумовые характеристики устройства сбора данных для обнаружения подозрительных областей на видео, записанном со статической сцены позволяют обнаружить подделку.

- Обнаружение на основе артефактов кодирования.

  На производительность метода обнаружения подделки видео в значительной степени влияет процесс кодирования видео. Кодирование видео выполняется для внедрения в видео артефактов, которые можно использовать для определения целостности видеоконтента путем его извлечения и проверки. В последние годы криминалисты в области видеонаблюдения полагаются на эти артефакты, чтобы определить целостность видео и найти в нем поврежденные области.

- Обнаружение несоответствий по содержанию.

  В конкретном видео сложно покадрово определить, являются ли геометрические, физические или световые свойства сцены настоящими и не поддельными. До сих пор были предложены два подхода для определения этого типа фальсификации:

  1) подход, основанный на артефактах, которые остаются в результате рисования видео

  2) подход, который раскрывает несоответствия в движении объектов

- Обнаружение копирования-перемещения в видео.

  Атаки подделки копирования-перемещения может быть выполнена на видео и разделена на два этапа, которые включают подделку перемещения внутрикадрового копирования и подделку перемещения межкадрового копирования. При внутрикадровой подделке часть кадра копируется и вставляется в другом месте этого кадра. Обычно это делается для того, чтобы скрыть или дублировать какой-либо объект одного кадра или нескольких кадров. При межкадровой подделке злонамеренные модификации применяются к последовательности кадров и различные типы этой атаки включают вставку кадра, удаление кадра, переупорядочивание кадров.

4.3. Общая структура обнаружения

Структура общего процесса обнаружения фальсификации видео состоит из 6 этапов. Первый шаг - разделить входное видео и извлечь из него кадры. Затем следует этап «Извлечение признаков» для поиска или извлечения векторов признаков. На этом этапе используются различные методы извлечения признаков, такие как DCT (дискретное косинусное преобразование), DWT (дискретное волновое преобразование) и т.д. На следующем этапе могут быть применены методы сопоставления перекрывающихся блоков, такие как дерево K-SVD (K-сингулярное разложение) и сортировка по основанию. После выполнения сопоставления блоков выполняются операции постобработки, и на последнем решающем этапе делается вывод о типе фальсификации видео и его местонахождении в кадре.

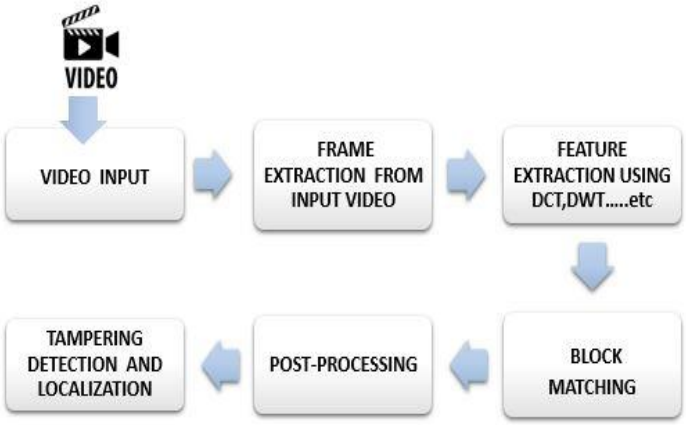

Рис. 15: Общая схема процесса обнаружения фальсификации

4.4. Сложные сценарии для идентификации видео

В некоторых системах наблюдения важными вопросами являются затраты на хранение и передачу данных. Для снижения затрат на хранение и передачу необходимо отправлять и хранить только те видеоклипы, которые содержат интересующие объекты. Более того, в большинстве приложений наблюдения фоновые объекты изменяются очень медленно по сравнению с объектами переднего плана. Возможное эффективное решение в этих сценариях заключается в том, что только объекты, представляющие интерес (в основном объекты переднего плана), отправляются кадр за кадром в режиме реального времени, в то время как фоновый объект отправляется один раз в течение длительного промежутка времени. В таких приложениях наблюдения становится очень важно защитить подлинность видео: подлинность от вредоносных изменений и подлинность для идентификации источника передачи (т. е. идентификации источника видео). В системах наблюдения, основанных на событиях, видеопоследовательность захватывается, когда происходит какое-либо изменение в сцене (наличие события) который будет запечатлен камерой. Если в сцене нет изменений, то камера наблюдения не захватывает никакой видеоряд. Этот вид системы наблюдения используется в военной системе для обеспечения безопасности границ. Аутентичность для такого рода видеопоследовательностей является сложной проблемой, поскольку в видеопоследовательностях, снятых камерой наблюдения, нет надлежащей временной последовательности. Это сценарии, которые создают значительные проблемы при проверке подлинности.

5. Дипфейки

Впервые распространение дипфейки получили в 2017 году, когда пользователь с ником "Deepfakes"на сайт Reddit начал размещать видео созданные с использованием алгоритма смены лиц на основе глубоких нейронных сетей Deep Neural Networks (DNN). Впоследствии термин DeepFake стал использоваться более широко для обозначения любых видео, в которых происходит замена лиц с помощью ИИ. Сейчас выделяют 3 основных вида видео DeepFake:

1. Синтез видео головы и верхней части плеча нужного человека с исходным. Поведение сохраняется

2. Замена лиц включает в себя создание нужное видео с лицами, замененными синтезированными лицами источника, сохраняя выражения лиц.

3. Синхронизация губ. Манипулируя губами заставлять говорить то, чего человек на самом деле не произносил.

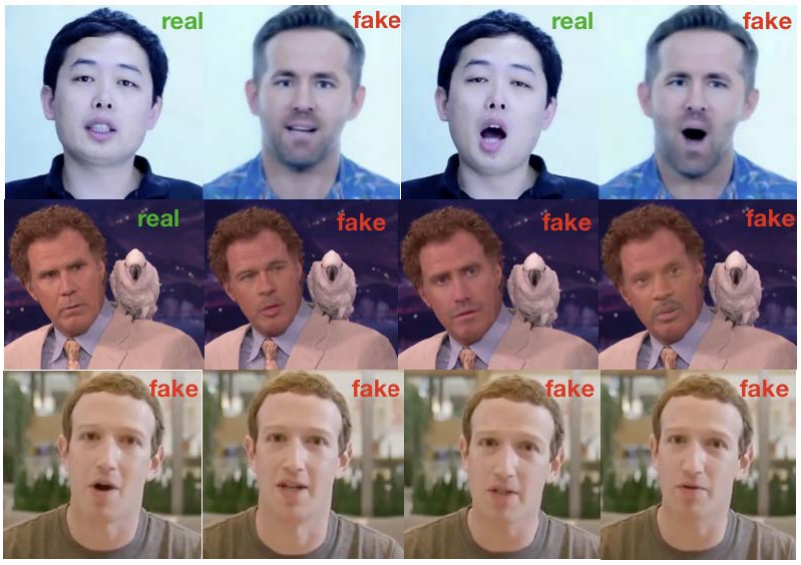

Рис. 16: Примеры DeepFake: 1 строка - 1 вид (синтез видео головы и плечей), 2 строка - 2 вид (замена лица с сохранением выражения), 3 строка - 3 вид (синхронизация губ)

Хотя есть интересные и творческие применения видео DeepFake, из-за сильной ассоциации лиц с личностью человека их также можно использовать в качестве оружия. Хорошо

составленные видеоролики DeepFake могут создавать иллюзии присутствия и действий человека, которые не происходят в реальности, что может привести к серьезным политическим, социальным, финансовым и юридическим последствиям.

5.1. Классификация методов обнаружения DeepFake

Текущие методы обнаружения DeepFake в основном нацелены на видео со сменой лиц, которые составляют большую часть видео DeepFake, распространяемых в Интернете. Многие из существующих методов сформулированы как задачи двоичной классификации на уровне кадра. В зависимости от используемых функций эти методы делятся на три основные категории, которые основаны на:

1. Физических несоответствиях.

   Во многих дипфейках отсутствует разумное моргание глаз из-за использования онлайн-портретов в качестве обучающих данных, которые обычно не закрываются по эстетическим причинам.

2. Артефактах.

   Визуальные артефакты видео DeepFake в существующих наборах данных, включая некачественные видимые границы сращивания, несоответствие цветов, видимые части исходного лица и несовместимые ориентации лица.

3. Управляемых данных.

   Используют различные типы глубоких нейронных сетей, обученные на реальных видео и фейковых видео, но фиксирующие определенные артефакты.

5.2. Сдерживание увеличения количества дипфейков

Так как датасет лиц для обучения алгоритмов машинного обучения берут из сети, и очень часто используют в недобросовестных целях, то один из предлагаемых вариантов сдерживать рост количества дипфейков, собираемых в веб-пространстве - метод нарушения синтеза лиц ИИ. Метод основан на добавлении возмущения к исходному изображению, чтобы отвлечь детекторы лиц на основе DNN, так что качество полученного набора лиц в качестве обучающих данных снижается.

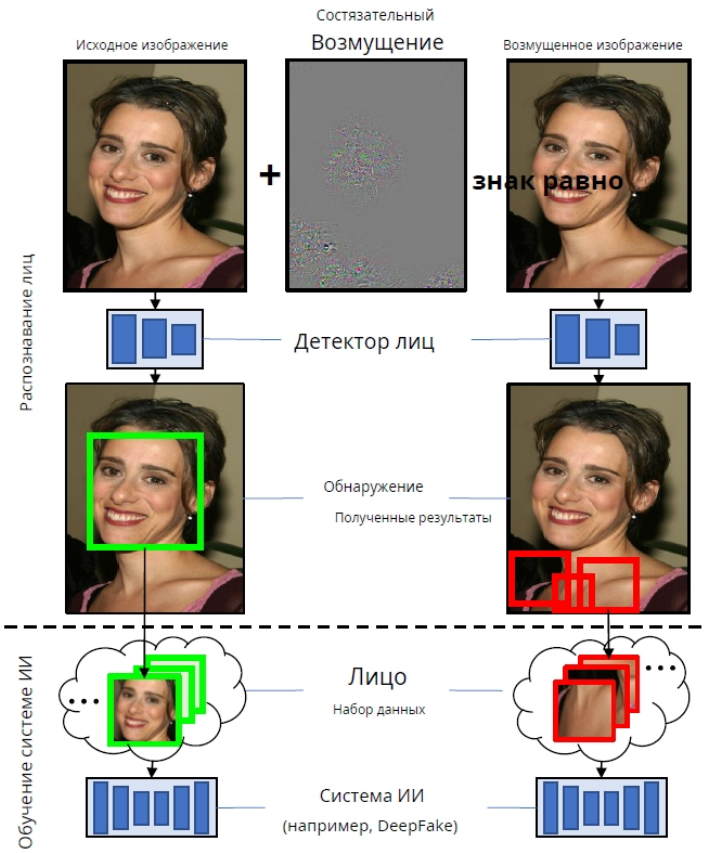

Рис. 17: Схема метода нарушения синтеза лиц ИИ

5.3. Дикие дипфейки

На данный момент в интернет выгружено достаточно большое количество дипфейков, и так как фейковая информация в этих видео потенциально создает большую информационную опасность, поэтому важно обнаруживать дипфейки. Но для обучения детекторов дипфейков требуется большой датасет из реальных и дипфейк видео. Обычные виртуальные дипфейки содержат сцены с маленьким количеством людей и похожие выражения лица, чаще всего просто разговор. А вот WildDeepFake (дикие дипфейки) сцены могут иметь огромное количество человек, более 10. А также, если обычные виртуальные дипфейки создаются с помощью популярных методов, то дикие дипфейки могут использовать различные комбинации методов и их создание более сложное. Виртуальные дипфейки чаще всего имеет низкое качество, в области лица могут быть заметные искажения, размытость и другие странные артефакты. Большинство диких дипфейков намеренно создаются в высоком качестве. В [4] был создан датасет диких дипфейков (WildDeepFake), состоящих

из 707 хорошо сделанных дипфейк-видео из интернета, видеоконтент разнообразен: множество действий (например, трансляция, фильмы, интервью, беседы и многое другое), разнообразные сцены, фоны и условия освещения, а также различные степени сжатия, разрешения и форматы.

5.4. Генерация дипфейков

Одним из широко используемых методов глубокого обучения для генерации дипфейков являются генеративные состязательные сети (GAN). На GitHub существует множество программ для дипфейков с открытым исходным кодом, например Faceswap-GAN и Faceswap. Большинство этих программ для дипфейков используют архитектуру кодировщика-декодера с одним кодировщиком и двумя декодерами: кодировщик изучает общие черты исходного (реального) и целевого (поддельного) лица, в то время как два декодера обучаются отдельно генерировать исходное и целевое лицо. Во время процесса замены лица декодер, связанный с исходным лицом, принимает кодировку целевого лица и генерирует поддельное исходное лицо. Маска внимания исходного лица обычно используется для того, чтобы поддельное лицо источника выглядело более убедительно с помощью этапа слияния. Сгенерированные поддельные лица могут быть дополнительно улучшены за счет использования изображений лиц с более высоким разрешением (как исходных, так и целевых).

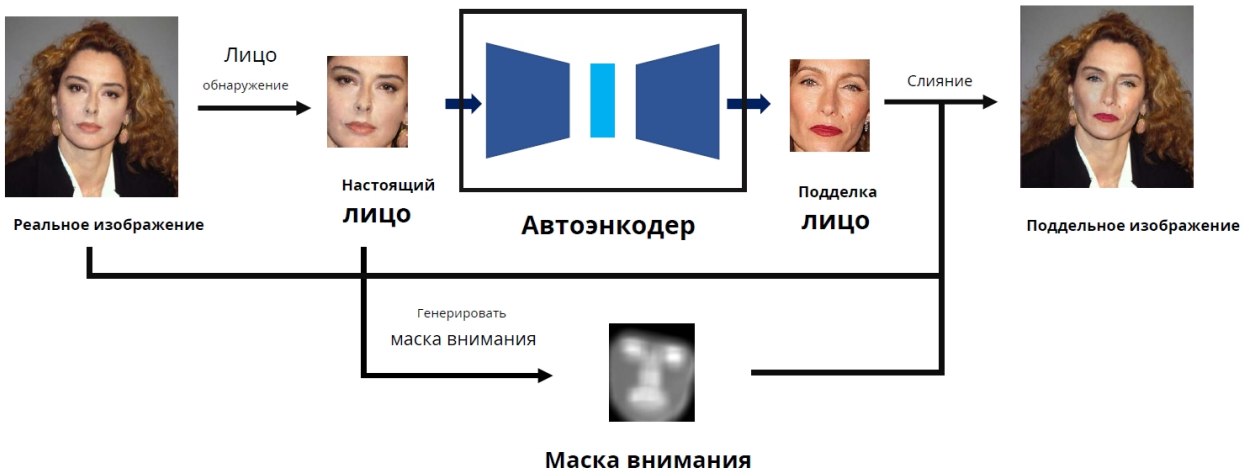

Рис. 18: Процесс замены лиц

5.5. Сеть обнаружения Deepfake

Одной из эффективных сетей обнаружения Deepfake - сеть на основе внимания (ADDNets), которая также может достаточно эффективно определять дикие дипфейки. Берется набор данных $D = (x, y)^(i)$ где $i = 1, .., n$ с $x \in X \subset R^{F*W*H*C}$ и $y \in Y = [0, 1]$ обозначающе видео и метку класса. F, W, H, C - количество кадров, ширина кадра, высота кадра и цветовой канал соответственно. Два класса, $y = 0$ реальные видео и $y = 1$ дипфейки. Цель обнаружения дипфейка – обучить бинарный классификатор f, который отображает пространство видео в пространство классов $f : X- > Y$. Это может достигаться за счет минимизации ошибки классификации f при обучении с данных D: $\underset{\theta}{\text{argmin}}\text{E}_{(\text{x},\text{y})\in\text{D}}\text{l}(\text{f}(\text{x}), \text{y})$, где l-функция потерь, θ - обучаемые параметры сети.Входные видео (настоящие и дипфейки) обрабатываются сначала извлекая лица и затем можно использовать сеть для обучения распознавания. Напрямую нельзя тренировать, т.к дипфейки изменяют только область лица.

5.6. Схема работы ADDNets для видеопоследовательностей

Сеть принимает входные данные последовательностей лиц. Длина последовательности L. Для каждого изображения лица в последовательности создается его маска внимания, используя один и тот же модуль генерации маски внимании. Каждая пара изображений: лицо и его маска внимания обрабатывается отдельным блоком добавления, которые имеют одинаковые веса (масштабируется входная маска в соответствии с выходным разрешением). Выходной слой сверточной нейросети имеет два нейрона соответствующие двум классам (реальный или фейк). По итогу входное видео соответствует одному из двух классов: реальные видео или фейк.

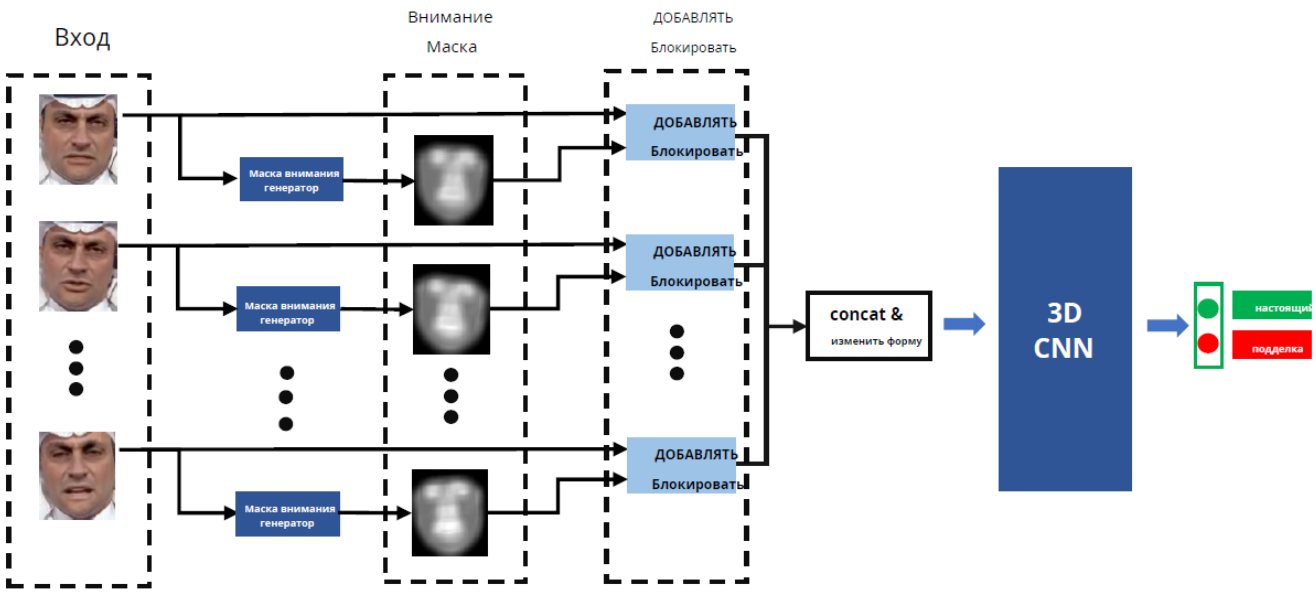

Рис. 19: ADDNets 3D для обнаружения дипфейк подделок на уровне видеопоследовательностей

5.7. Результаты обнаружения с помощью ADDNets

Network	DFD	DF-TIMIT LQ	DF-TIMIT HQ	FF++(Deepfake) LQ	FF++(Deepfake) HQ	Wild-Deepfake
P3D	70.16%	76.71%	62.25%	67.05%	75.23%	53.20%
C3D	73.18%	94.44%	82.38%	87.72%	95.00%	55.87%
I3D	67.83%	**96.38%**	**89.85%**	**93.18%**	96.70%	62.69%
ADDNet-3D	**94.93%**	90.17%	85.75%	90.11%	**98.30%**	**65.50%**

Рис. 20: Сравнение уровня точности распознавания дипфейков с помощью ADDNet в зависимости от датасета

Рассматривались три существующих наборов данных DeepfakeDetection (DFD), DeepFake-Timit (DF-TIMIT) и FaceForensics++ (FF++), для последних двух было две версии , как низкое качество LQ, так и высокое качество HQ. , а также созданный датасет WildDeepfake. Точность обнаружения дипфейков – производительность метрики. Если с существующими наборами данных DFD, TIMIT, F++ точность распознавания доходит от 90 до 95%, то с дикими дипфейками никакие базовые сети не могу достичь точности выше 70%. Среди базовых сетей I3D обеспечивает лучшую производительность почти для всех

протестированных наборов данных, за исключением DFD, где C3D более эффективен. В частности, I3D обеспечивает высокую точность обнаружения$>$ 89% для четырех наборов данных DF-TIMIT и FF ++, тогда как C3D имеет точность 73,18% для DFD. Опять же, все они значительно падают на WildDeepfake с гораздо более низкой точностью $<$63%. Но ADDNet обеспечивает удивительно высокую точность 94,93% на наборе DFD, не сильно отстает на наборах DF-TIMIT В FF++, но заметно лучше распознает WildDeepFake, чем остальные сети.

Исходя из [4] получаются разные результаты успеха верного вычисления дипфейков в зависимости от самих датасетов. И меньше всего точность распознавания у диких дипфейков, которые является более сложным набором данных, где производительность базовых детекторов может резко снизиться, т.к они состоят из лиц в движении, которых в кадре может быть несколько, а также изначально высокое качество видео. Также стоит отметить, что по сравнению с 2D-сетями обнаружения, 3D-сети, как правило, менее эффективны. Одной из возможных причин снижения производительности сетей трехмерного обнаружения является то, что временная информация, содержащаяся в последовательностях глубоких подделок лиц, также искажается из-за покадровой генерации поддельных лиц. Это указывает на то, что временная информация в дипфейке видео следует обрабатывать иначе, чем в реальном видео, чтобы повысить точность обнаружения глубоких подделок на уровне последовательности.

6. Заключение

В данной работе были рассмотрены разные виды атак на видеопоследовательности и некоторые методы обнаружения этих атак.

1. Были исследованы атаки с пространственным вмешательством, временным вмешательством, пространственно-временным вмешательством и дипфейки.

2. Были проанализированы активные и пассивные методы обнаружения, сеть обнаружения дипфейков на основе внимания ADDNets.

Необходимо продолжать исследовать атаки на видеопоследовательности и еще больше совершенствовать методы для их обнаружения.

Список литературы

[1] S. Upadhyay, S.K. Singh. Video Authentication: Issues and Challenges, 2012.

[2] L.C.Manikandan, R. Habeeb. Video Tampering Attacks and Detection Techniques,2019

[3] S. Lyu. DeepFake Detection: Current Challenges and Next Steps, 2020.

[4] B. Zi, M. Chang, J. Chen, X. Ma, Y. Jiang. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection, 2021.

[5] N. Aggarwal, R.Singh. Video content authentication techniques: a comprehensive survey, 2017

[6] S. Upadhyay, S.K. Singh, M. Vatsa, and R. Singh. Video authentication using relative correlation information and SVM, 2008.

[7] A.Gandhi, S.Jain. Adversarial Perturbations Fool Deepfake Detectors, 2020.

[8] N. Ruiz, S.A. Bargal,S. Sclarof.Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems, 2020.

[9] P. Neekhara, B. Dolhansky, J. Bitton, C.C Ferrer. Adversarial Threats to DeepFake Detection: A Practical Perspective, 2020.

[10] P. Yin, H. Yu, Classification of Video Tampering: Methods and Countermeasures using Digital Watermarking, 2001

[11] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, B.-G. Kim. Deepfake Detection Scheme Based on Vision Transformer and Distillation, 2021.

[12] A. Gironi, M. Fontani, T. Bianchi, A. Piva, M. Barni. A video forensic technique for detecting frame deletion and insertion, 2014.

[13] Stamm, C. Matthew, W. Sabrina Lin, KJ Ray Liu. Temporal forensics and anti-forensics for motion compensated video, 2012.

[14] N. Mondaini, R. Caldelli, A. Piva, M. Barni, and V.Cappellin. Detection of malevolent changes in digital video for forensic applications, 2007.

[15] V. Amanipour, S. Ghaemmaghami. Video-Tampering Detection and Content Reconstruction via Self-Embedding, 2018.

[16] O.I. Al-Sanjary, G. Sulong. Detection of video forgery: a review of literature, 2015.

[17] J. Hussein, A. Mohammed. Robust Video Watermarking using Multi-Band Wavelet Transform, 2009.

[18] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection, 2020.

[19] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer. The DeepFake Detection Challenge (DFDC) Dataset.

[20] S. Fung, X. Lu, C. Zhang, C.-T. Li. DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning, 2021.