Новый DeepFake метод замены лица на видео на основе удаления фона

А.В. Кузнецов[1,2]

[1]*Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001*
[2]*Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34а, Самара, Россия, 443086*

Аннотация

Современное цифровое пространство ежедневно насыщается огромным объёмом данных в виде изображений и видео. Вся содержащаяся информация представляет важность для пользователей, организаций и других её потребителей. Следует отметить, что скорость распространения информации настолько высока, что порой для того, чтобы отследить её первоисточник, необходимо затратить большие вычислительные ресурсы. Более того, в ходе поиска первоисточника, информация может исказиться до неузнаваемости и приобрести новые свойства в виде деталей и дополнительных связей. Злоумышленники изменяют информацию, как правило, в целях компрометации личности, разжигания конфликтов, подрыва репутации физических и юридических лиц и т.д. Среди наиболее современных и популярных средств является технология DeepFake, с помощью которой на изображении или видео может быть произведена замена лица человека. В работе предлагается новый метод генерации искажённых изображений, основанный на удалении фона на изображении, для повышения визуального качества искажения.

Ключевые слова

Искусственное искажение, deep fake, подделка, генеративно-состязательная сеть, удаление фона

1. Введение

Термин «информационная гигиена» всё чаще начинает использоваться в современном мире в связи с широким распространением ложной информации, которая намеренно искажает отражаемые в ней события. С ростом дезинформации неизбежно растёт и развивается аппарат защиты от распространения изменённой информации. Следует отметить, что, когда мы употребляем термин «ложная информация», мы можем подразумевать один из нескольких вариантов её формирования – от искажения существующих фактов путём добавления или удаления каких-либо существенных деталей до искусственной генерации новой информации, которая в реальной среде просто не имела места. В данной работе предлагается новый DeepFake метод создания искажённых изображений на основе удаления фона, что позволит создавать искажения более высокого визуального качества. В дальнейшем результаты генерации таких искажений будут использоваться для разработки метода их обнаружения.

2. Способы искажения цифровых изображений и видео

Существует большой набор методов и средств для внесения искусственных искажений в цифровые изображения и видео [1-3]. Стремительное развитие технологий искусственного интеллекта привело к появлению алгоритмов формирования подделок такого высокого уровня, который не может обнаружить даже профессиональный эксперт. Одной из таких современных технологий создания подделок является DeepFake [1]. Методы и алгоритмы, относящиеся к данной технологии, построены на использовании генеративных моделей машинного обучения

[2]. Созданные при помощи таких методов искажённые фото и видео данные характеризуются очень высоким уровнем качества. С предметной точки зрения алгоритмы DeepFake представляют из себя способы замены лиц на изображениях и видео. В настоящее время в открытых источниках существует большое количество видео примеров применения технологии DeepFake для замены лиц знаменитых актёров, например, замена лица Маколея Калкина на лицо Сильвестра Сталлоне в фильме «Один дома». Все эти примеры лишним образом подтверждают огромный спектр угроз, которые могут возникнуть при распространении такого рода ложной информации посредством сети Интернет и СМИ. Замена лица бывшего президента США лицом актёра Уила Смита была произведена при помощи метода, основанного на сегментации лица [3].

Рисунок 1: Пример применения технологии DeepFake

3. DeepFake метод искажения изображения с использованием устранения фона

В отличие от известного метода замены лица на видео [3], в основе которого лежит сегментация лица на 5, 10 и 15 сегментов и дальнейшее обучение свёрточной генеративно-состязательной сети, в данной работе предлагается использовать устранение фона и выделение объекта переднего плана для устранения артефактов сегментации на границе изменяемой области (именно выявление краевых эффектов). Это позволит избавиться от изменений, возникающих в смежных с заменяемой областью фрагментах. В ходе разработки предлагается новая структура генеративно-состязательной сети с автоэнкодером, который позволит учитывать фон при генерации искажённых изображений и, тем самым, снижать значение выходной функции потерь, что позволит повысить визуальное качество искажённых изображений.

4. Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 20-37-70053, 19-07-00138, 19-07-00474.

5. Литература

[1] Tolosana, R. Deepfakes and beyond: A survey of face manipulation and fake detection / R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia // ArXiv preprint: 2001.00179. – 2020.

[2] Goodfellow, I. Generative adversarial nets / I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio // Advances in Neural Information Processing Systems. – 2014. – Vol. 3(11). – P. 2672-2680.

[3] Siarohin, A. Motion Supervised co-part Segmentation / A. Siarohin, S. Roy, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe // arXiv preprint: 2004.03234. – 2020.