**СЛАЙД 1**

Добрый день, уважаемые участники конференции!

Я студент группы 150501 Климович Алексей Николаевич.

Я подготовил доклад на тему «АНИМАЦИЯ ИЗОБРАЖЕНИЙ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ» под руководством старшего преподавателя кафедры ЭВМ Куприяновой Дианы Вячеславовны.

**СЛАЙД 2**

Современные цифровые технологии оказывают заметное влияние на визуальный контент, трансформируя его в различных прикладных сферах – от кинематографа и видеоигр до AR/VR-приложений и социальных медиа.

Одной из основных задач в этой области является реалистичная анимация статичных изображений, позволяющая «оживлять» портреты, создавать цифровые аватары или автоматизировать производство мультимедийного контента.

Традиционные методы, основанные на ручной анимации или других алгоритмах, требуют значительных временных затрат и не обеспечивают естественности движений. Также для данного подхода требуются дополнительные сведения об объекте анимации, например, его 3D-модель, необходимая для более корректной генерации движения, что стимулирует развитие нейросетевых методов, способных обходиться без явного 3D-представления.

**СЛАЙД 3**

С учетом этого в данной работе были выдвинуты следующие цели:

* + проведение сравнительного анализа современных методов анимации изображений на основе нейронных сетей и оценка их качества;
  + реализация и обучение модели нейронной сети для анимации изображений;
  + тестирование модели и сбор метрик ее работы.

**СЛАЙД 4**

На сегодняшний день существуют несколько архитектур нейронных сетей, которые способны решать задачу анимации изображений по заданному видео.

Во-первых, это проект X2Face, представленный в 2018 году некоторыми исследователями из Оксфорда и компании Visual Geometry Group. Это один из первых проектов, который смог доказать, что в принципе анимация изображений на основе нейронных сетей возможна!

Она позволяет управлять изображением лица, используя:

* Видео;
* Другие изображения;
* Аудио или код позы.

Вся модель обучается self-supervised методом, то есть без ключевых точек, 3D-моделей или размеченных данных.

Во-вторых, это Monkey-Net – модель, представленная в 2019 году группой людей из итальянского университета в сотрудничестве с компанией Snap Inc.

Monkey-Net – первая модель, которая предлагает универсальный способ анимации изображений любых объектов (не только лиц), основанный на обучаемых ключевых точках и модели движения, получаемой напрямую из видео. Она также работает в self-supervised режиме – не требует разметки или заранее заданных keypoints.

В-третьих, проект 2020 года – First Order Motion Model – это, можно сказать, доработка предыдущего проекта теми же людьми.

Главное отличие от предыдущих подходов (в частности, Monkey-Net) – это использование аффинных преобразований при моделировании движения, что дает более точный и качественный результат.

**СЛАЙД 5**

На данном слайде представлено небольшое сравнение перечисленных моделей.

Как видно, архитектура модели, как и модель движения, усложняется в хронологическом порядке.

Также стоит отметить, что в двух последних моделях есть поддержка маски окклюзии (occlusion map), что немаловажно, но об этом я расскажу чуть попозже.

**СЛАЙД 6**

Далее идет продолжение сравнения.

Как видно из таблицы, более сложная архитектура модели в целом улучшила качество генерации анимации, но при этом ухудшилась сама скорость генерации. Поэтому самой быстрой сетью оказалась X2Face, а самой медленной – First Order Motion Model.

Также на данном слайде выделены основные плюсы и минусы сравниваемых сетей.

Например, несмотря на то, что X2Face проигрывает в качестве генерации, у нее есть возможность генерации изображения не только по видео, но и по другим входным данным, что описывается свойством мульти-модальности.

**СЛАЙД 7**

В конце концов, для реализации и тестирования была выбрана архитектура FOMM, которая изображена на данном слайде.

Как видно, на вход подается исходное изображение и управляющее видео. Далее из исходного изображения и кадров видео извлекаются ключевые точки (по отдельности), а также вычисляются аффинные преобразования. После этого, эти данные вместе с исходным изображением подаются в модуль плотного поля движения, где вычисляется само поле движения и маска окклюзии.

Поле движения – это карта пиксельных перемещений от исходного изображения к управляющему кадру, используя:

* Разность позиций ключевых точек;
* Аффиные преобразования;
* Heatmaps (тепловые карты), описывающие влияние каждой ключевой точки.

Маска окклюзии показывает, какие области были закрыты/открыты между кадрами.

Далее эта информация и исходное изображение подается в модуль генерации, который уже и занимается генерированием новых кадров для результирующего видео.

**СЛАЙД 8**

На данном слайде представлены технологии, используемые для реализации и тестирования данной модели.

Во-первых, это язык программирования Python и фреймворк PyTorch, которые стали очень популярными и удобными для работы в области Машинного Обучения, Компьютерного зрения и т.д.

Для обучения использовалась среда Google Colab, так как компания Google любезно предоставляет бесплатное использование их серверов с довольно неплохими видеокартами для работы с нейронными сетями на 3 часа в день, что довольно сильно ускоряет разработку и обучение. Также для локального тестирования использовался PyCharm, как одна из самых подходящих сред программирования на языке Python.

В качестве набора данных использовался VoxCeleb – набор данных, который представляет собой видеопоследовательности интервью с лицами людей знаменитостей, извлеченных из 22496 видео YouTube.

**СЛАЙД 9**

После обучения, были собраны некоторые метрики, отражающие качество моделей и которые представлены в следующей таблице.

1. L1 – среднее расстояние между сгенерированным и реальным видео. Измеряется в диапазоне от 0 до 1, где 0 – идеальное совпадение по пикселям.
2. AKD (Average Keypoint Distance) – cреднее расстояние до ключевых точек. Ключевые точки вычисляются независимо для каждого кадра. AKD получается путем вычисления среднего расстояния между обнаруженными ключевыми точками реального видео и сгенерированного видео. Значения AKD выражаются в пикселях.
3. AED (Average Euclidean Distance) – среднее евклидово расстояние между истинным и сгенерированным представлением кадра. AED – мера, показывающая семантическую схожесть изображений. Чем меньше ее значение, тем более «узнаваемым» и реалистичным является результат.

**СЛАЙД 10**

На данном слайде представлены полученные результаты анимирования изображений по видео.

Слева расположено исходное изображение, сверху даны кадры ведущего видео. Также сами результаты ранжированы по их моделям так, чтобы можно было проанализировать и сравнить выходные изображения (кадры видео).

Также я прикрепил гифку с анимацией портрета Моно Лизы по видео с моим лицом, чтобы наглядно продемонстрировать, как это работает.

**СЛАЙД 11**

На данном слайде приведены используемые источники информации для данного проекта.

Это, во-первых, описание приведенных сетей – X2Face, Monkey-Net и First Order Motion Model, а также описание выбранного датасета для обучения модели.

**СЛАЙД 12**

На этом всё, спасибо за внимание!