ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 004.932.4

Отчет об исследовательском проекте на тему:

Улучшение качества изображения с помощью методов компьютерного зрения

Выполнил студент:

группы #БПМИ223, 3 курса Волянский Никита Денисович

Принял руководитель проекта:

Лукьянченко Петр Павлович
Старший преподаватель
Факультет компьютерных наук НИУ ВШЭ

Москва 2025

Содержание

Аннотация

Данная работа посвящена реализации модели, позволяющей повышать качество и привлекательность изображений людей в конкретной одежде посредством генерации изображения с новым человеков в той же позе и той же одежде с возможным изменением фона, с использованием методов компьютерного зрения. Основное внимание в работе уделяется изучению современных методов детекции позы человека на изображении, генерации изображения нового человека по заданной позе, сегментации участков тела, одежды и фона, маскирования одежды и переноса маски с одного человека на другого, генерации и изменения фона изображения, а также их реализации и применении в новой модели. В результате были проанализированы существующие методы для задач, описанных выше, выбраны наиболее подходящие для решаемой проблемы и получена модель, комбинирующая эти методы и позволяющая решать прикладную задачу улучшения качества и повышения привлекательности изображения одежды на человеке.

Ключевые слова: Глубинное обучение, компьютерное зрение, сегментация, генерация изображения, трансформеры, диффузионные модели

1 Введение

В наши дни концепция онлайн-шоппинга развивается стремительными шагами, все больше и больше людей предпочитают сделать заказ онлайн вместо того, чтобы тратить свое свободное время на долгое пребывание в оффлайн-магазинах. На фоне этого многие представители малого бизнеса переходят в онлайн, пользуясь, в том числе, возможностями маркетплейсов. Однако у тех предпринимателей, которые занимаются продажей одежды, может возникать проблема отсутствия бюджета или возможностей для проведения качественной фотосессии со своим ассортиментом, ведь для этого требуется нанять модель, фотографа, снять студию и пр. и даже при наличии всего вышеперечисленного нет гарантии в получении желаемого результата. Тем не менее, качественные фотографии ассортимента необходимы как продавцам для привлечения новых покупателей, так и самим покупателям, которые могут упустить вещь, возможно заинтересующую их. Одним из решений этой проблемы может послужить модель, принимающая на вход изображение человека, не обязательно высокого качества, в нужной одежде, генерирующая нового человека и перенося на него эту одежду, меняющая фон на более презентабельный, повышающая разрешение и выдающая на выходе улучшенное изображение с новым человеком на привлекательном

фоне, но в той же позе и одежде, что был человек на начальном изображении. Данный метод поможет предпринимателям приблизится к качеству студийной фотографии, при этом не потратив много ресурсов.

1.1 Постановка задачи

Задачей этого проекта является изучение, сравнение и подбор релевантных методов, наболее подходящих под каждую подзадачу модели, нахождение их подходящей комбинации, внедрение в новую модель и реализация итоговой полной модели, решающей поставленную выше проблему. Можно выделить следующие подзадачи:

- Распознание и классификация позы, в которой стоит человек
- Классификация человека на изображении по полу и возрасту (если эти параметры не заданы заранее)
- Генерация нового человека в распознанной позе
- Сегментация одежды, человека и заднего фона на первоначальном изображении
- Плавный перенос одежды на сгенерированного человека, чтобы одежда сидела на нем естественно
- Генерация и добавление на фотографию заднего фона
- Повышение разрешения конечного изображения

Реализация и тестирование разных версий модулей и модели в целом будет проводиться на языке Python с использованием инструментов PyTorch, зарекомендовавших себя одними из самых удобных и прикладных в сфере глубинного обучения и компьютерного зрения в частности.

1.2 Актуальность

Данная работа актуальна, так как в сегодняшнем мире современных технологий, в особенности после пандемии, популярность онлайн-шоппинга заметно выросла, но применение нейронных сетей в этой сфере еще не очень широко. Добавление возможности улучшения качества и привлекательности изображения товаров одежды даст развитие отрасли покупки и продажи элементов одежды, принося пользу как продавцам, так и покупателям.

2 Обзор литературы

2.1 Определение положения человека

Первая задача, с которой мы сталкиваемся - это определение позиции человека на фотографии. Под позицией в первом случае будем понимать набор соединенных ключевых точек, соответсвующих суставам, шее, голове. Чаще всего в контексте нашей задачи нужно будет определять положение только одного человека, но будем рассматривать методы также позволяющие распознавать положения и нескольких людей на одном изображении.

DeepPose [1] - первое значимое решение, использующее методы глубинного обучения. В данном методе задача определения позиции человека рассматривается как задача совместной регрессии по ключевым точкам и решается с помощью сверточных нейронных сетей глубокого обучения. Модель основана на AlexNet и дополнительном слое, выводящим координаты ключевых точек. При обучении используется L2-loss для регрессии. Преимущества данной модели заключаются в том, что неросеть для каждой ключевой точки рассматривает весь ее контекст, а также проще формулируется в отличие от других моделей.

Более современная модель DeepCut [13] преимущественно отличается от предшествующей тем, что в ней возможно определение позиций для нескольких людей на одном изображении, но большого прогресса в задаче определения позиции для одного человека не произошло

Наиболее релевантной можно считать модель OpenPose [4].

OpenPose [4] была первой системой, решающей задачу определения позиций для нескольких людей в режиме реального времени и продемонстрировавшей возможность распознавания свыше 100 ключевых точек на одном изображении. OpenPose [4] поддерживает определение не только крупных частей, но и отдельных пальцев и их движений. Работа в OpenPose [4] происходит следующим образом:

- Изображение передается в сверточную нейронную сеть для получения предсказаний ключевых точек
- Затем создается карта достоверности для идентификации частей человеческого тела на изображении
- Генерируются поля сходства фрагментов (PAFs) для кодирования попарных отношений между частями тела
- Затем части тела, принадлежащие одному человеку, связываются с помощью двустороннего сопоставления

- Каждая часть объединяется в единую позицию для каждого человека на изображении

В ходе работы над проектом появилась возможность рассмотреть позицию человека немного в другом смысле - а именно в виде очертаний человека, его конечностей, лицевых признаков и в том числе одежды на нем.

Наиболее релевантным алгоритмом для поиска границ на изображении является Canny edge detector [3], включающий следующие шаги:

- Применение фильтра Гаусса для сглаживания изображения с целью удаления шума
- Вычисление градиентов изображения, границы отмечаются в максимальных значениях градиента
- Подавление немаксимумов. Границами отмечаются только локальные максимумы
- Применение двойной пороговой фильтрации для определения возможных границ
- Подавление всех слабых границ, не связанных с сильными

2.2 Определение пола и возраста человека

В том случае, если на вход модели подается только изображение, а это самый удобный вариант для рядового пользователя, нужно уметь определять пол и возраст человека на изображении для более точной генерации нового человека.

Наиболее релевантным и удобным для использования фреймворком является DeepFace [17]. Это гибридная система распознавания лиц, объединяющая state-of-the-art модели, такие как FaceNet [16], OpenFace [2], DeepFace [5] и др. Модель, отвечающая за классификацию человека по полу достигает 97.44% accuracy, что является вполне хорошим результатом для решения нашей подзадачи.

2.3 Генерация изображения человека в заданной позе и заднего фона

Для генерации фона по текстовому описанию подойдет Stable Diffusion [14] - text-to-image модель глубокого обучения, основанная на методах диффузии. На сегодняшний день одна из самых используемых диффузионных моделей для генерации изображений по текстовому запросу.

В подзадаче генерации человека помимо самого запроса есть дополнительное условие - заданная поза, в которой должен стоять человек, поэтому просто Stable Diffusion [14] здесь не обойтись. В недавней работе ControlNet [21] как раз рассматривался способ управления процессом генерации изображений в Stable Diffusion [14] с помощью дополнительных условий, одним из которых может быть позиция человека, которую можно передать вместе с запросом. В таком случае модель будет гененрировать изображение не только по текстовому описанию, но также будет опираться на переданные условия. Таким образом, для генерации человека в заданной позе можно воспользоваться ControlNet [21].

2.4 Сегментация изображения

Для сегментации изображения существуют различные известные модели, показывающие хорошие результаты в этой сфере, например, U-Net [15], Mask R-CNN [9], Fully convolutional networks [12].

В нашей задаче особое внимаение нужно уделить именно сегментации одежды, частей тела человека, не покрытых одеждой, и заднему фону, поэтому нужно рассмотреть более специализированные модели. Хорошие результаты в этой области показывают модели SCHP(Self-Correction for Human Parcing) [10] и SegFormer [19], дообученный на датасете ATR [11]

2.5 Перенос одежды на сгенерированного человека

Для переноса одежды можно рассмотреть два подхода:

1 Брать изображение сгенерированного нового человека и изображение вырезанной одежды с оригинала и пытаться выполнить «виртуальную примерку одежды», т.е. пытаться наложить одежду со второго изображения на человека с первого так, чтобы сохранялась текстура и одежда сидела естественным образом

2 Наложить изображение вырезанной с оригинала одежды поверх изображения сгенерированного человека в исходном виде и дальше убирать неровности и несостыковки в некоторых местах

В качестве решений для первого подхода можно рассматривать модели виртуальной примерки одежды, такие как VITON [8], VITON-HD [6] и аналоги. Их общая идея - имеются фотография человека, на которого нужно перенести одежду и изображение элемента одежды с видом спереди. Выделяются тело и ключевые зоны человека, определяется поза,

затем одежда деформируется по позе человека и с помощью модуля примерки переносится естественным образом на человека. Поэтому такие модели могут быть одним из решений переноса одежды.

Также для первого подхода можно рассмотреть идею IP-Adapter [20], когда диффузионная модель генерирует новое изображение по промпту в виде изображения, а не текста. С применением технологии Inpaining для IP-Adapter [20] можно попробовать передавать в качестве image-prompt фотографию одежды, маскировать область, где должна находиться эта одежда на сгенерированном изображении с новым человеком и подавать на вход модели.

Для второго подхода можно рассматривать различные варианты подавления шума на изображении либо же можно перегенерировать изображение так, чтобы оно не сильно поменялось, но некоторые несоответствия ушли. Подходящим примером для перегенерации изображения является Stable Diffusion [14], но уже в Image-to-image формате. Можно подобрать наиболее подходящее значение гиперпарамтера, влияющего на степень изменения изображения так, чтобы максимально пытаться сохранить текстуру оригинальной одежды и при этом сгладить неровности.

2.6 Повышение разрешения конечного изображения

Так как для быстрой и качественной генерации Stable Diffusion [14] требует изображения небольшого размера (обычно 512x512) и на выходе дает изображение такого же размера, то нужно дополнительно последним шагом увеличить разрешение перегенерированной картинки. Учитывая, что полученные после генерации изображения не сильно зашумленные, то для этого можно использовать модель RealESRGAN [18], которая не требует больших вычислительных ресурсов, но отлично справляется с задачей.

3 Выбор составляющих для итоговой модели

В разделе с обзором литературы были рассмотрены различные способы решения некоторых подзадач модели и теперь нужно выбрать те из них, которые, во-первых, будут показывать хорошие результаты обособленно от итоговой модели для своих подзадач, и, во-вторых, будут хорошо комбинироваться с другими частями итоговой модели.

(a) Original image (b) OpenPose result

Рис. 3.1: Пример сложной позы

3.1 Определение положения человека

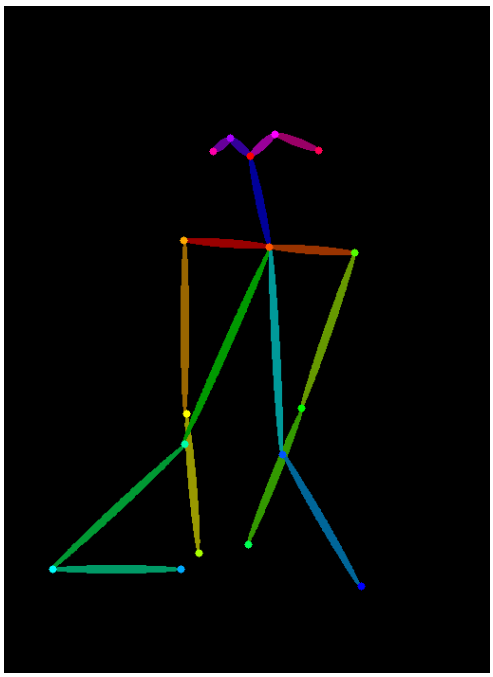
OpenPose [4]

Сначала я решил попробовать OpenPose в качестве условия, которое передается ControlNet и на которое опирается Stable Diffusion при генерации изображения. Сам OpenPose определяет позицию человека довольно точно даже в сложных случаях, как показано на Рисунке 3.1.

Существенные недостатки этого способа в контексте нашей задачи проявляются уже при генерации изображения с использованием полученной позы как условия:

- Для сложных поз диффузионная модель не всегда может понять точно, как и какие конечности располагаются друг относительно друга. Пример неудачной генерации можно видеть на Рисунке 3.2. Диффузионная модель не смогла понять положение руки и ноги относительно друг друга, предполагаемо из-за того, что на полученной позе конечности пересекаются, в том числе в точках сочленения

- Диффузионная модель ограничивается только позой человека, понимаемой как набор соединенных ключевых точек, поэтому при генерации не обязательно будет сохранено телосложение человека с оригинальной фотографии, волосы (что важно, если на оригинальном изображении на человеке есть головной убор), аксессуары. Также диффузионная модель хоть и сохраняет общую концепцию переданной позы, но может



(a) OpenPose result

(b) ControlNet + Stable Diffusion result

Рис. 3.2: Пример неудачной генерации по сложной позе

интерпретировать ее по разному в разных случаях. Пример можно увидеть на Рисунке 3.3, видно что поза одна, но диффузионная модель расположила конечности немного по-разному, волосы тоже лежат по-разному.

- Самое важное, что в диффузионную модель не передается никакой информации об изначальной одежде человека и поэтому при генерации модель ничем не ограничена в этом аспекте, но это гораздо сильнее усложняет задачу переноса оригинальной одежды на сгенерированного человека, потому что есть вероятность того, что на нового человека придется не только "надевать"одежду с изначального изображения, но и "снимать"лишнюю сгенерированную (например, если на человеке с оригинального изображения были шорты, а после генерации только по позе диффузионная модель выдала человека в брюках). На Рисунке 3.3 видно, что даже на двух примерах одежда отличается, на правом примере рукава длинные, на левом укороченные, а на изначальном изображении их нет. Также нет и шляпы, которая была на оригинальном изображении с Рисунка 3.1

Canny edge detector [3]

Учитывая все минусы использования OpenPose было принято решение использовать Canny edge detector, так как полученные с помощью этого алгоритма границы человека на

(a) Generation result 1

(b) Generation result 2

Рис. 3.3: Пример двух генераций по одной позе

изображении сохраняют не только информацию о позе, понимаемой как набор соединенных ключевых точек и получаемой из OpenPose, но также и об одежде, ее форме, головных уборах и аксессуарах. Также этот способ гораздо сильнее ограничивает диффузионную модель в положении человека на сгенерированной фотографии и расположении его конечностей, потому что есть четкие границы, в которые нового сгенерированного человека нужно вписать. Пример границ, получаемых с помощью алгоритма Canny edge detector и результатов генерации по этим границам можно увидеть на Рисунке 3.4. Заметно, что в отличие от использования OpenPose положение сгенерированного человека практически в точности повторяет положение оригинального, вне зависимости от конкретной генерации. Также сохраняется информация о форме одежды и шляпы.

По полученным результатом было принято решение использовать в дальнейшем генерацию с использованием Canny edge detector [3]

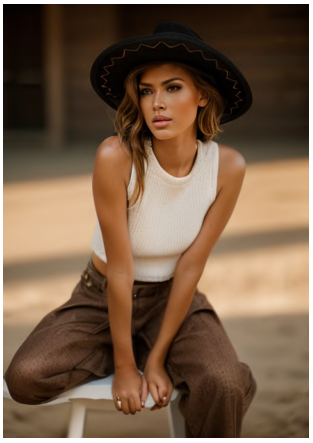
3.2 Сегментация изображения

Это довольно важный этап, потому что именно на нем мы определяем на оригинальном изображении одежду, которую хотим перенести и качество итогового результата напрямую зависит от того, насколько хорошо модель для сегментации умеет отличать элементы одежды, головные уборы, аксессуары от открытых частей тела человека и от заднего фона.

Необходимо рассмотреть и сравнить две упомянутые ранее модели:

(a) Original image

(b) Canny edge detector result



(c) Generation result 1

(d) Generation result 2

Рис. 3.4: Получение границ человека с помощью Canny Edge detector и результаты генерации

1 Self-Correction for Human Parcing [10], которая была обучена на Human Parcing датасетах и должна хорошо отделять части тела человека и одежды

2 SegFormer [19], дообученный на Human Parcing датасете ATR [11] и тоже должен справляться с поставленной задачей

На Рисунке 3.5 приведены некторые из примеров сегментации. По ним можно заметить, что SCHP очень чувствителен к изменению оттенков, потому что пытается не только отделить одежду от человека и от заднего фона, но еще и максимально точно отделить разные элементы одежды друг от друга, и из-за этого возникают неточности, которые сильно повлияют на итоговый результат в итоговой модели. Например, на первом изображении часть правого рукава была принята за задний фон, на втором изображении есть артефакты на лице и туловище, на третьем немного обрезано лицо. Так как в дальнейшем одежда будет вырезаться по полученным маскам на этапе сегментации, такое частое наличие неточностей будет приводить к не очень качественному итоговому результату.

Если рассматривать результаты, полученные с помощью дообученного SegFormer, то можно увидеть, что он не так хорошо отличает друг от друга разные элементы одежды, как SCHP, зато он гораздо точнее отделяет в целом одежду от частей тела человека и заднего фона, а это намного важнее в нашей задаче.

По полученным результатом было принято решение делать сегментацию начального изображение с помощью SegFormer [19], дообученного на Human Parcing датасете ATR [11].

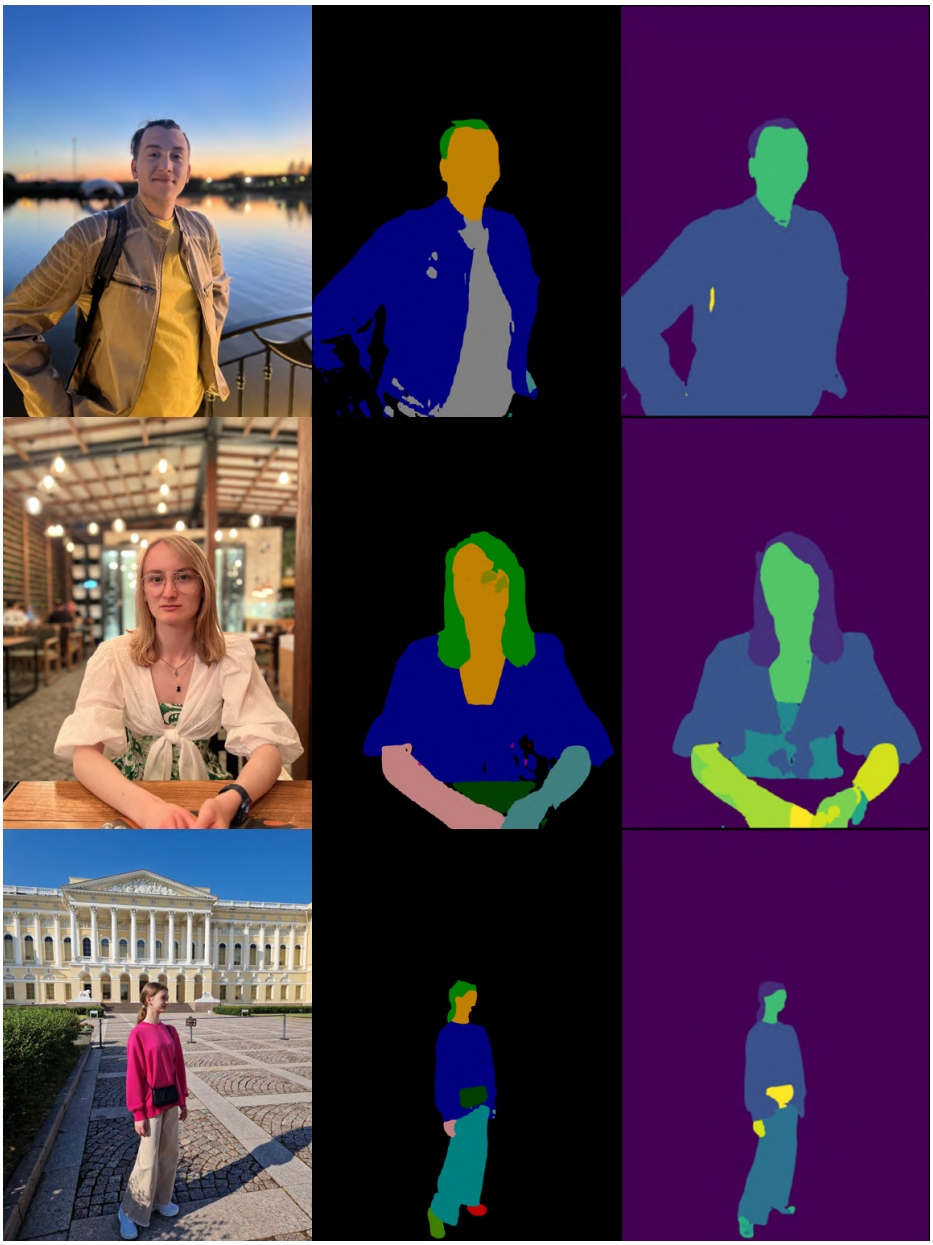


Рис. 3.5: Примеры сегментации. Слева - оригинальное изображение, по центру - результат SCHP, справа - результат fine-tuned SegFormer

3.3 Перенос одежды на сгенерированного человека

Первый подход

Как упомяналось ранее, идея первого подхода заключается в том, чтобы имея сгенерированное изображение нового человека и вырезанную с изначального изображения одежду после сегментации, пытаться перенести одежду на человека так, чтобы она сохраняла текстуру и сидела естественно. Для решения этой задачи было предложено несколько способов. Вариант с моделями для виртуальной примерки одежды не подошел, так как обычно на вход они принимают один элемент одежды с видом спереди и должны точно знать к какой части тела должен этот элемент относиться. Также они не могут переносить головные уборы, аксессуары, обувь. Попытки передать одежду целиком, полученную после сегментации, не увенчались успехом, модели не могли понять куда ее переносить.

Другой предложенный подход - использование IP-Adapter [20] для закрашивания (Inpainting сгенерированного изображения по маске. На вход подавались сгенерированное ранее изображение нового человека и одежды. По человеку определялась маска туловища для закрашивания, а одежда использовалась в качестве image-prompt для диффузионной модели. Полученные результаты не оправдали ожиданий, модель не смогла понять как правильно перенести всю одежду. Результаты можно увидеть на Рисунке 3.6

Второй подход

Идеей же второго подхода было наложение вырезанной одежды поверх сгенерированного изображения и дальнейшее редактирование неровностей. Так как в модели используется Canny edge detector при генерации нового человека, то на полученном изображении позиция сгенерированного человека получается практически идентичной оригиналу, а также сохраняется форма одежды. В силу этого при простом наложении одежды поверх сгенерированного изображения получается довольно неплохой результат, за исключением того, что есть некторые неровности по краям одежды из-за неидеального сегментирования и в некоторых местах небольшие несостыковки в местах соприкосновения одежды и открытых частей тела человека. Также, скорее всего, нужно будет подкорректировать освещение и тени на одежде, чтобы они соответствовали той среде, в которой находится сгенерированный человек.

Первым решением для этой задачи могли послужить алгоритмы и модели, уменьшающие шум на изображении, производящие сглаживание. Но такой подход приводил к тому, что размывалось все изображение, а не только неровные края. Также они никак не меняли освещение и тени. Некоторые результаты, полученные таким путем можно увидеть на

(a) Generated image input

(b) Cut clothes input

(c) IP-Adapter result 1

(d) IP-Adapter result 2

Рис. 3.6: Результаты работы IP-Adapter

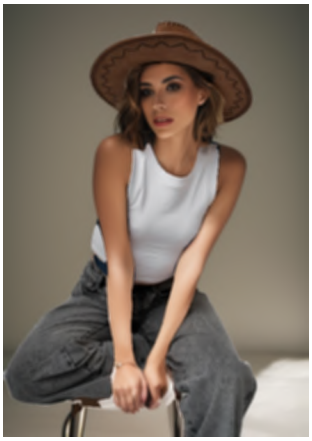
Рисунке 3.7.

Вторым решением было предложено использовать Stable Diffusion [14] в Image-to-Image формате, который зашумляет переданное изображение и расшумляет его, опираясь на переданный промпт. Некоторые результаты его работы можно увидеть на Рисунках 3.8 3.9. Видно, что данная модель хорошо справляется со сглаживанием контуров одежды, добавлением плавных переходов между открытыми частями тела и одеждой и немного изменяет одежду по форме, чтобы она сидела более естественно на сгенерированном человеке.

В итоге результаты, полученные с помощью Stable Diffusion [14] в Image-to-Image формате оказались наиболее качественными и дальше в модели будет использоваться этот подход.

Теперь, после того, как для подзадач были выбраны наиболее подходящие решения, можно собрать все в одну итоговую модель.

(a) Generated image + cut clothes (Input)



(b) Gauss filter

(c) NlMeans denoising

Рис. 3.7: Примеры уменьшения шума

(a) Generated image + cut clothes

(b) Regenerated image

(c) Generated image + cut clothes (zoomed)

(d) Regenerated image (zoomed)

Рис. 3.8: Сравнение изображений до и после Image-to-image генерации Stable Diffusion

(a) Generated image + cut clothes

(b) Regenerated image

(c) Generated image + cut clothes (zoomed)

(d) Regenerated image (zoomed)

Рис. 3.9: Сравнение изображений до и после Image-to-image генерации Stable Diffusion

4 Описание модели

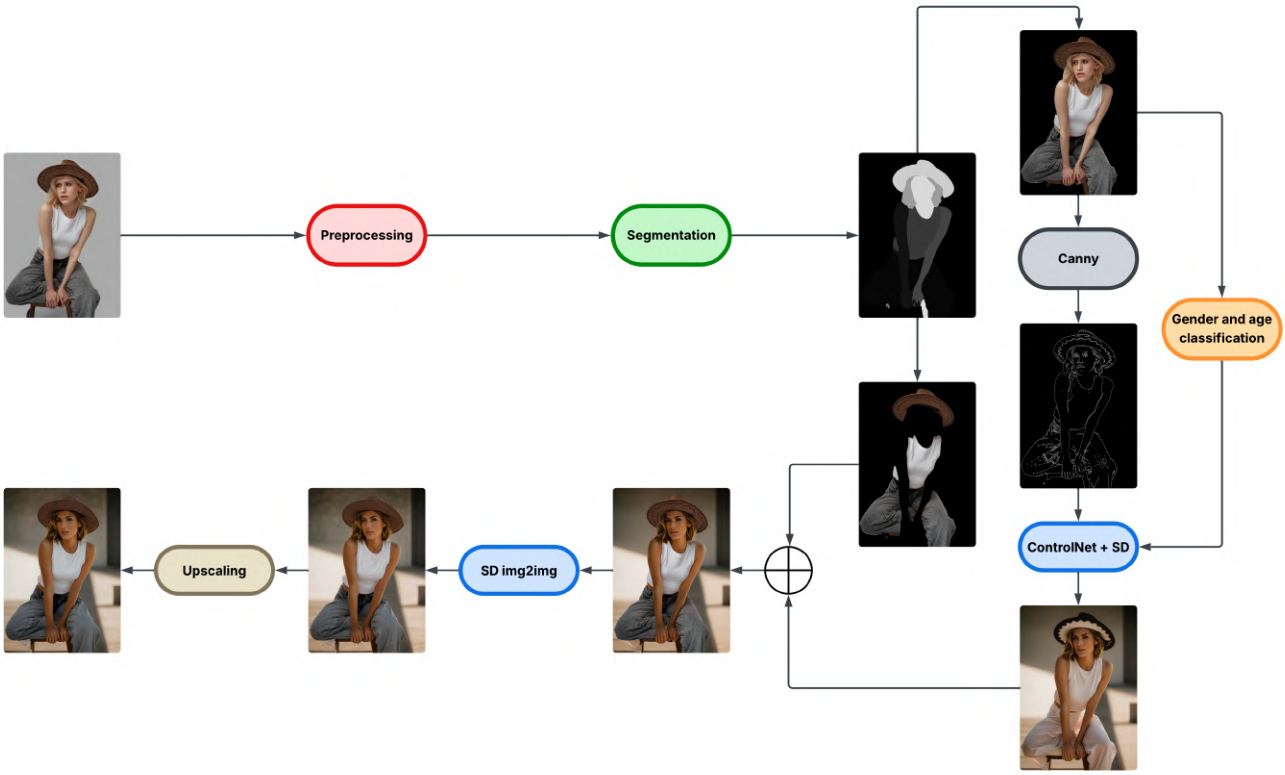


Рис. 4.1: Иллюстрация работы модели

Диаграмму работы модели можно увидеть на Рисунке 4.1

4.1 Предобработка изображения

Так как Stable Diffusion долго генерирует большие изображения и так как на выходе она дает изображения, высота и ширина которых кратна 8, а в нашей задаче важно работать с изображениями одного размера, то требуется предварительная обработка изображения.

На этапе предобработки сначала максимум из высоты и ширины изображения приводится к 1024, а минимум соответсвенно умножается на коэффициент $\frac{1024}{max(height,width)}$. Далее изображение обрезается симметрично так, чтобы отнормированный минимум из высоты и ширины был кратен 8.

4.2 Сегментация изображения

На этапе сегментации производится сегментация предобработанного изображения с помощью SegFormer [19], дообученного на Human Parcing датасете ATR [11]. Далее из полученных предсказаний вычисляются маска для изображения человека без заднего фона и

маска для всей одежды. По полученным маскам модель вырезает человека и одежду соответственно и получает два новых изображения.

4.3 Классификация пола и возраста человека

Полученное на предыдущем шаге изображение человека без заднего фона подается на вход DeepFace [17], который анализирует лицо человека и предсказывает для него пол и возраст.

4.4 Canny edge detector

По полученному изображению человека без заднего фона строятся границы с помощью алгоритма Canny edge detector [3].

Для выделения границ берется только изображение человека, потому что мы также хотим иметь возможность менять задний фон на изображении и тогда для генерации не нужна информация об объектах, находящихся рядом и позади человека на оригинальной фотографии

4.5 Генерация изображения по заданным границам

Получив информацию от классификатора можно составить текстовый промпт для Stable Diffusion. Также у пользователя есть возможность задать желаемый задний фон, информация о котором тоже добавляется в промпт.

Далее, используя полученный конкретизированный промпт, полученные на предыдущем шаге границы человека на изображении, применяется ControlNet со Stable Diffusion для генерации нового человека в той же позиции, что и на оригинальном изображении и на новом фоне.

4.6 Наложение изображений и перегенерация результата

На этом шаге изображение одежды, полученное на шаге сегментации, накладывается поверх сгенерированного изображения, полученного на прошлом шаге и итоговый результат подается на вход Image-to-Image Stable Diffusion модели с тем же текстовым промптом, что был на прошлом шаге. На выходе получаем изображение со сглаженными краями, подкорректированным освещением и более реалистичным наложением одежды на сгенерированного человека.

4.7 Повышение разрешения

На последнем шаге перегенерированное изображение, полученное на прошлом шаге подается на вход модели RealESRGAN [18] и апскейлится в 4 раза. На выходе этой модели и модели в целом получаем изображение нового сгенерированного человека на новом фоне, но в такой же позе и одежде, что человек на оригинальном изображении.

5 Эксперименты и подбор гиперпараметров

Парамтеры low_threshold и high_threshold для Canny edge detector

Эти параметры в алгоритме Canny edge detector влияют на фильтрацию границ - чем меньше эти параметры, тем более мелкие границы захватывает алгоритм и полученные очертания становятся более детализированными.

Если брать их слишком маленькими, например 25 и 75 соотвественно, то резульатом алгоритма будут слишком детализированные границы и периодически даже лишние, на которые диффузионная модель будет обращать внимание при генерации. Например, на Рисунке 5.1 видно, что модель дорисовала ожерелье, которого не было на начальном изображении, но за которую диффузионная модель посчитала одну из мелких границ. В целом видно, что диффузионная модель пытается максимально детализировать генерируемое изображение, но в контексте нашей задачи это излишне.

Рис. 5.1: Результат работы Canny edge detector и модели в целом при low_threshold = 25, high_threshold = 75

При больших значениях парамтеров, напрмер 200 и 250 соотвественно, границы наоборот получаются очень недетализированными и даже некоторые основные из них, которые

показывают границы конечностей, прерываются. Это тоже плохо для модели, потому что мы меньше начинаем ограничивать диффузионную модель в позиции человека и она может добалять некоторые детали от себя. Пример можно видеть на Рисунке 5.2

Рис. 5.2: Результат работы Canny edge detector и модели в целом при low_threshold = 200, high_threshold = 250

Оптимальнее всего оказалось брать значения посередине между двумя такими границами и я использую low_threshold = 100 и high_threshold = 150. С ними границы получаются достаточно детализированными, чтобы ограничивать диффузионную модель в позе человека и форме одежды на нем, но в то же время не слишком сильно детализированными, чтобы они могли быть перепутаны диффузионной моделью с чем-то другим.

Оптимальное количество шагов для генерации изображения по границам

Нужно выбрать оптимальное количество шагов для генерации изображения в ControlNet + Stable Diffusion, чтобы качество сгенерированного изображения было высоким, оно выглядело реалистично, но при этом генерировалось не слишком долго. При выборе колчества шагов, не превосходящего 20-25 (в зависимости от того, насколько много места лицо знаимает в кадре) лицо плохо генерировалось, были артефакты, несоответствие цвету. Стабильно качественные результаты возникали при выборе количества шагов, начиная с 30, поэтому я решил взять этот параметр равным 40, чтобы даже на изображениях, на которых лицо занимает мало пикселей, генерация лица была приемлемой, без лишних артефактов. При выборе такого количества шагов генерация изображения по границам из Canny edge detector занимает около 35-40 секунд, что не сильно долго для качественного результата. Результаты для малого и выбранного количества шагов показаны на Рисунке 5.3

(a) Original image (b) Result, steps = 15 (c) Result, steps = 40

Рис. 5.3: Результаты работы модели при генерации изображения по границам с разным числом шагов

Параметр strength для Image-to-Image Stable Diffusion

Это довольно значимый параметр в контексте нашей задачи, потому что нужно найти баланс между тем, чтобы сохранять уникальность одежды, ее текстуру, детали, но при этом сгладить неровности, образующиеся при накладывании изображения одежды на сгенерированное изображение человека.

При выборе параметра около 0.1 изображение практически не меняется и, соотвественно, все неровности остаются. Но уже при значении около 0.35 диффузионная модель хорошо сглаживает все неровности, делает освещение и тени на одеждя более естественными для среды, в которой находится сгенерированный человек, но начинает сильно менять детали на одежде, например, надписи, логотипы, рисунки и т.п.

После подбора значений наиболее сбалансированным вариантом оказалось значение strength = 0.25, при котором модель практически всегда (иногда с небольшими помарками) сглаживала изображение и преобразовывала одежду так, чтобы она сидела естественным образом на сгенерированном человеке и при этом не слишком сильно меняла изначальную текстуру и детали на одежде. Примеры результатов для разных значений парамерта strength можно видеть на Рисунке 5.4

Эксперименты с разными видами изображений

В модель могут передавать изображения различных видов, например, в полный рост, по пояс, далеко от камеры, и нужно проверить как модель будет с ними работать

(a) Original image

(b) Result, strength = 0.1

(c) Result, strength = 0.25

(d) Result, strength = 0.35

Рис. 5.4: Результаты работы модели при перегенерации наложенных изображений с разным параметром strength

В полный рост, человек не занимает большую часть изображения

Пример работы можно видеть на Рисунке 5.5. Видно, что когда человек не занимает большую часть изображения, диффузионной модели отводится меньше пикселей для генерации лица, и она не всегда генерирует его хорошо. Тем не менее одежда перенесена качественно.

Фотография по пояс

Пример работы можно видеть на Рисунке 5.6. С таким вариантом фотографии лицо уже занимает больше пикселей на изображении и после генерации оно получается более детализированным, чем в предыдущем случае. Одежда все так же хорошо перенеслась и узор на пальто стал даже более ярковыраженным.

(a) Original image

(b) Result

Рис. 5.5: Пример результата, когда на оригинальном изображении человек занимает меньшую часть фотографии

Много лишних объектов на фотографии

Пример работы можно видеть на Рисунке 5.7. Видно, что модель довольно хорошо справилась с отделением человека от заднего фона и перенесла все аксессуары, но при этом посчитала сиденье велосипеда частью одежды.

Эксперименты с задним фоном

На вход модели можно также передать информацию о желаемом заднем фоне на новой фотографии и на Рисунке 5.8 представлены некоторые примеры для одной входной фотографии и разный уточнений насчет заднего фона. Видно, что с простыми и довольно абстрактными примерами, такими как «forest», «ocean», «New-York» (Нью-Йорк можно просто считать мегаполисом) модель справляется довольно хорошо. Если же подавать на вход что-то более специфическое (как «Kremlin»), то модель не всегда сможет точно такой задний фон воссоздать.

6 Заключение

В процессе работы над проектом было изучено большое количество моделей и алгоритмов для разпознания позы человека на изображении, получения границ объектов на изображении, сегментации изображения и специализированно направленных на сегментацию одежды, генерации изображений, в том числе с использованием некоторых условий,

(a) Original image (b) Result

Рис. 5.6: Пример результата работы модели с фотографией по пояс

(a) Original image (b) Result

Рис. 5.7: Пример результата работы модели с фотографией с большим количеством объектов

виртуальной примерки одежды и сглаживания неровностей на изображении. Были найдены наиболее подходящие модели, как в смысле решения подставленных подзадач, так и в смысле встраиваимости в общую модель. В результате была получена новая модель для улучшения качества и привлекательности изображений людей в конкретной одежде, аналогов которой на сегодняшний день нет. Результаты работы выложены в репозитории на GitHub [7].

Можно выделить некоторые преимущества полученной модели:

- Работает с обычными фотографиями, которые каждый может сделать на камеру своего телефона и в результате дает новую фотографию студийного качества в высоком разрешении с красивым фоном и отретушированной одеждой
- Может работать с фотографиями разного размера, на которых человек изображен как в полный рост, так и по пояс

(a) background = "Forest"

(b) background = "New-York"

(c) background = "Ocean"

(d) background = "Kremlin"

Рис. 5.8: Пример результатов работы модели на одном входном изображении с разным параметром заднего фона

- Можно дополнительно задавать в параметрах желаемый задний фон, а также пол и возраст

Также у модели можно выделить некоторые недостатки:

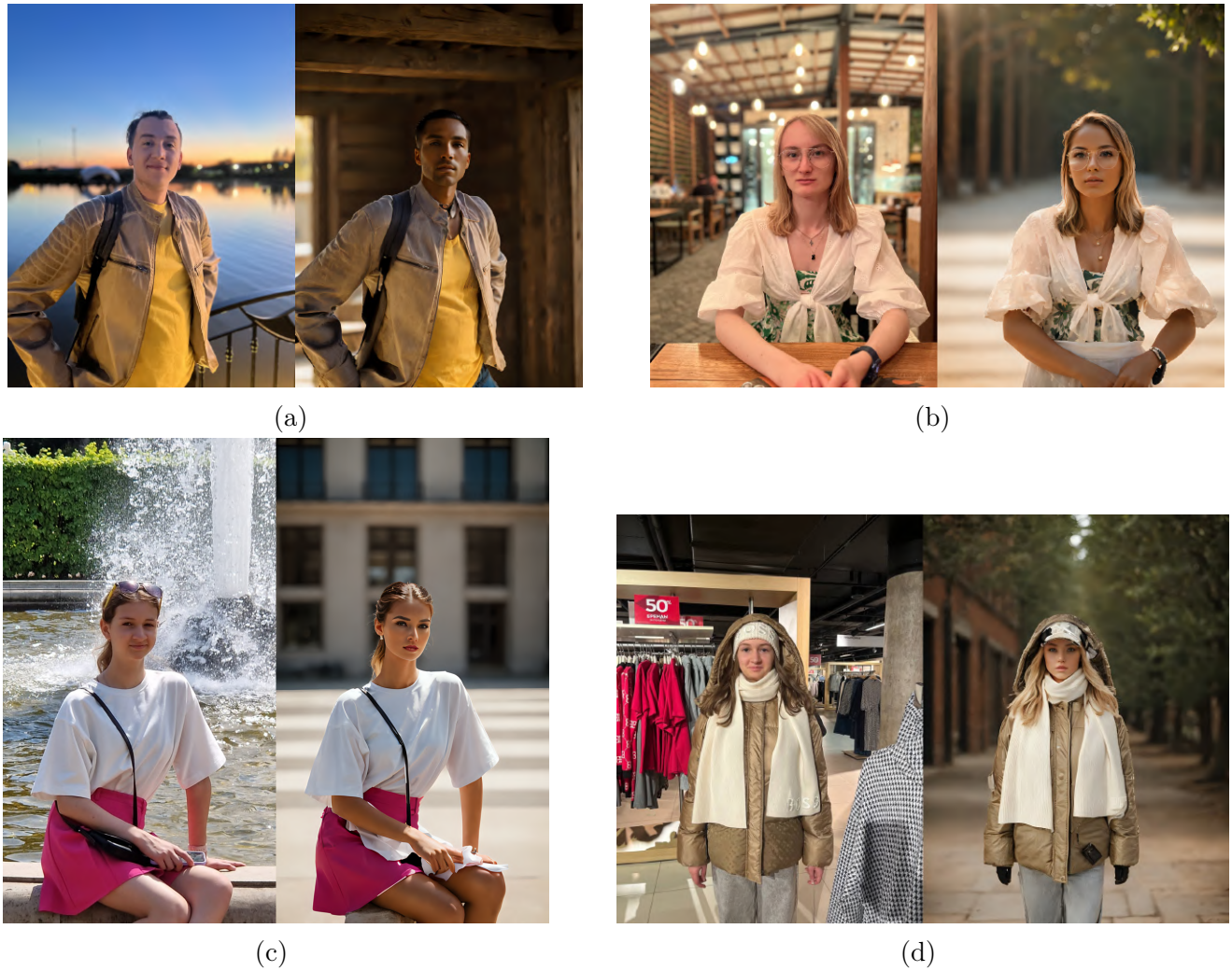
- Сегментация одежды не всегда работает на 100% точно. Если на изображении есть небольшие просветы между частями тела, например, рукой и туловищем, то модель для сегментации может посчитать этот просвет частью одежды. Также, если посторонние предметы находятся рядом с человеком, то они тоже могут быть приняты за часть одежды

- Если одежда имеет много мелких деталей, сложный узор, рисунок, то при прохож-

дении Image-to-Image Stable Diffusion блока, он может немного видоизменить их. Это то, чем жертвует модель для того, чтобы одежда сидела на сгенерированном человеке естественно

Поэтому в качестве перспективы дальнейшей деятельности можно пытаться улучшить точность сегментации одежды, которая будет учитывать мелкие просветы и лишние предметы на изображении, а также модифицировать действущий способ наложения одежды на сгенерированного человека или разработать новый, который будет меньше изменять надписи и сложные узоры на одежде.

Примеры работы модели можно посмотреть на Рисунке 6.1



(a) (b)

(c) (d)

Рис. 6.1: Примеры работы модели

Список литературы

[1] Christian Szegedy Alexander Toshev. “DeepPose: Human Pose Estimation via Deep Neural Networks”. В: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[2] Tadas Baltrušaitis, Peter Robinson и Louis-Philippe Morency. “OpenFace: An open source facial behavior analysis toolkit”. В: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, с. 1—10. DOI: 10.1109/WACV.2016.7477553.

[3] John Canny. “A Computational Approach To Edge Detection”. В: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* PAMI-8 (дек. 1986), с. 679—698. DOI: 10.1109/TPAMI.1986.4767851.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei и Y. A. Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. В: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[5] Hardie Cate, Fahim Dalvi и Zeshan Hussain. *DeepFace: Face Generation using Deep Learning*. 2017. arXiv: 1701.01876 [cs.CV]. URL: https://arxiv.org/abs/1701.01876.

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee и Jaegul Choo. *VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization*. 2021. arXiv: 2103.16874 [cs.CV]. URL: https://arxiv.org/abs/2103.16874.

[7] *GitHub link to repository with the model*. URL: https://github.com/NickVolya/Improving-Image-Quality-model (дата обр. 29.04.2025).

[8] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu и Larry S. Davis. *VITON: An Image-based Virtual Try-on Network*. 2018. arXiv: 1711.08447 [cs.CV]. URL: https://arxiv.org/abs/1711.08447.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár и Ross Girshick. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV]. URL: https://arxiv.org/abs/1703.06870.

[10] Peike Li, Yunqiu Xu, Yunchao Wei и Yi Yang. “Self-Correction for Human Parsing”. В: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). DOI: 10.1109/TPAMI.2020.3048039.

[11] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin и Shuicheng Yan. “Deep Human Parsing with Active Template Regression”. В: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37.12 (дек. 2015), с. 2402—2414. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2015.2408360.

[12] Jonathan Long, Evan Shelhamer и Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. 2015. arXiv: 1411.4038 [cs.CV]. URL: https://arxiv.org/abs/1411.4038.

[13] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler и Bernt Schiele. “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation”. В: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. URL: http://arxiv.org/abs/1511.06645.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser и Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

[15] Olaf Ronneberger, Philipp Fischer и Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[16] Florian Schroff, Dmitry Kalenichenko и James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. В: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, июнь 2015, с. 815—823. DOI: 10.1109/cvpr.2015.7298682. URL: http://dx.doi.org/10.1109/CVPR.2015.7298682.

[17] Sefik Serengil и Alper Ozpinar. “A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules”. В: *Journal of Information Technologies* 17.2 (2024), с. 95—107. DOI: 10.17671/gazibtd.1399077. URL: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077.

[18] Xintao Wang, Liangbin Xie, Chao Dong и Ying Shan. “Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data”. В: *International Conference on Computer Vision Workshops (ICCVW)*. 2021.

[19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez и Ping Luo. “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers”. В: *CoRR* abs/2105.15203 (2021). arXiv: 2105.15203. URL: https://arxiv.org/abs/2105.15203.

[20] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han и Wei Yang. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. 2023.

[21] Lvmin Zhang, Anyi Rao и Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023.