

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	6
1. Обзор моделей машинного обучения, связанных с задачей генерации изображения человека по текстовому описанию	8
1.1. Задача сегментации изображения.....	8
1.1.1. Модели для сегментации изображения человека.....	8
1.1.2. Задача детализированной сегментации изображения человека	10
1.2. Генеративно-состязательные сети	12
1.2.1. Условные генеративно-состязательные сети	13
1.2.2. Генеративно-состязательные сети с метрикой Вассерштейна для состязательной функции потерь	13
1.3. Модели для генерации изображения по текстовому описанию....	14
1.4. Наборы данных для генерации позы по текстовому описанию....	15
1.5. Модели для генерации изображения человека с измененной позой	16
Выводы по главе 1	16
2. Обзор предложенного решения задачи генерации изображения человека по текстовому описанию	18
2.1. Описание решения задачи генерации изображения человека по текстовому описанию	18
2.2. Описание решения для задачи детализированной сегментации изображения человека	19
2.3. Описание решения для задачи генерации позы по текстовому описанию	20
2.3.1. Набор данных для генерации позы по текстовому описанию	20
2.3.2. Модель для генерации позы по текстовому описанию	21
2.3.3. Вспомогательные функции потерь модели для генерации позы по текстовому описанию	22
2.3.4. Описание генератора позы по текстовому описанию	23
2.3.5. Описание дискриминатора модели для генерации позы по текстовому описанию	23
2.4. Описание решения для задачи генерации изображения с новой позой	25

2.5. Набор данных для алгоритма генерации позы по текстовому описанию	26
Выводы по главе 2	26
3. Обзор полученных результатов для задачи генерации изображения по текстовому описанию.....	28
3.1. Полученные результаты для задачи детализированной сегментации изображения человека	28
3.2. Полученные результаты для задачи генерации позы человека по текстовому описанию	28
3.3. Тестирование результатов для задачи генерации позы человека по текстовому описанию	29
Выводы по главе 3	30
ЗАКЛЮЧЕНИЕ	31
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	32

ВВЕДЕНИЕ

Компьютерное зрение — это направление в области искусственного интеллекта и связанные с ним технологии, которые позволяют извлекать значимую информацию из изображений, видео и других визуальных входов, чтобы впоследствии выполнять разные действия или строить рекомендации на основе этой информации.

Одной из популярных задач компьютерного зрения является автоматическое создание изображений на основе текста. Существует множество подходов к решению этой задачи, однако современные методы по прежнему не позволяют создавать реалистичные изображения на основе сложных подписей к изображениям из разнородной области.

Целью настоящей работы является разработка и реализация алгоритма генерации изображения человека на основе детализированной сегментации существующего изображения человека, соответствующее данному текстовому описанию.

Генеративно-состязательные сети [3] (англ. Generative Adversarial Networks, GANs), решающие эту задачу, способны генерировать нереалистичные изображения низкого качества. Это обусловлено отсутствием декомпозиции в подходах к решению данной задачи, в том смысле, что большинство моделей генерируют сразу изображение, которое будет соответствовать многим признакам, указанным в тексте, а также отсутствием качественного набора данных, на котором могли бы обучаться модели.

Данная работа предлагает гораздо более простую, но более эффективную модель генерации изображения человека в новой позе, соответствующей текстовому описанию, чем в предыдущих работах. В соответствии с указанными недостатками данный алгоритм имеет большое число преимуществ, среди которых:

- а) Декомпозиция в задаче генерации человека в новой позе, релевантной текстовому описанию;
- б) Улучшение результатов алгоритма, за счет проделанной детализированной сегментации.

Предлагаемый алгоритм принимает на вход изображение человека и текстовое описание на английском языке. Изначально осуществляется детализированная сегментация изображения человека, что позволяет подавать будущей

модели на вход изображения без внешнего шума. Для того, чтобы осуществить это качественно, было использовано матирование (англ. alpha matting) изображения, а также была сделана детализированная сегментация посредством выделения лица и одежды у человека. Следующим этапом является генерация позы по текстовому описанию с использованием порождающей состязательной сети. Последним этапом является генерация нового изображение с изменной позой, полученной шагом ранее, по данному изображению, обработанному с использованием детализированной сегментации.

Данный подход позволит получать реалистичные изображения людей с позой, соответствующей данному текстовому описанию, что может быть использовано в индустрии моды для интернет-магазинов, в игровой индустрии для выбора одежды для героев и т.д. К примеру, для пользователей интернет-магазинов одежды можно будет осуществить генерацию желаемой позы модели в понравившемся товаре.

В главе 1 проведен обзор существующих моделей для сегментации изображений человека, в том числе детализированной, генерации изображений по текстовому описанию и трансляции изображени с подробным описанием решений, которые будут использованы впоследствии.

В главе 2 предлагается детальное описание реализованного решения, с использованием технических выкладок и схем при необходимости.

В главе 3 проведен анализ решения: описаны результаты и проведена их оценка.

ГЛАВА 1. ОБЗОР МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ, СВЯЗАННЫХ С ЗАДАЧЕЙ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЯ ЧЕЛОВЕКА ПО ТЕКСТОВОМУ ОПИСАНИЮ

В этой главе будет произведен обзор общих понятий, связанных с поставленной задачей, а также будут рассмотрены текущие существующие модели машинного обучения и наборы данных, которые могут быть использованы для достижения цели проекта.

1.1. Задача сегментации изображения

Фундаментальной задачей компьютерного зрения является задача поиска групп пикселей, каждая из которых характеризует один смысловой объект, на изображениях и видео. Различают два типа подхода для данной задачи: сегментация и матирование.

Сегментация изображения — это задача кластеризации частей изображения на группы, посредством предсказания на уровне пикселей. То есть, сегментация изображения генерирует двоичное изображение, в котором пиксель либо принадлежит одной группе, либо другой.

Матирование изображения отличается от сегментации изображения тем, что некоторые пиксели могут принадлежать как одной группе, так и другой, такие пиксели называются частичными или смешанными пикселями. Чтобы полностью отделить передний план от фона в изображении, необходима точная оценка альфа-значений для частичных или смешанных пикселей.

Традиционные алгоритмы матирования изображений требуют для упрощения задачи трехканальные карты (англ. trimap) (при условии, что на входе одно изображение RGB), указывающих вероятность того, что каждый пиксель принадлежит каждому из трех классов (передний план, фон, неопределенная область). На рисунке 1 представлен пример трехканальной карты для изображения.

1.1.1. Модели для сегментации изображения человека

Матирование людей, высококачественное выделение людей на естественных изображениях с разными текстурами вне объекта, имеет решающее значение для самых разных приложений. Для данной работы выделение человека на изображении является важной частью, потому что в дальнейшем это

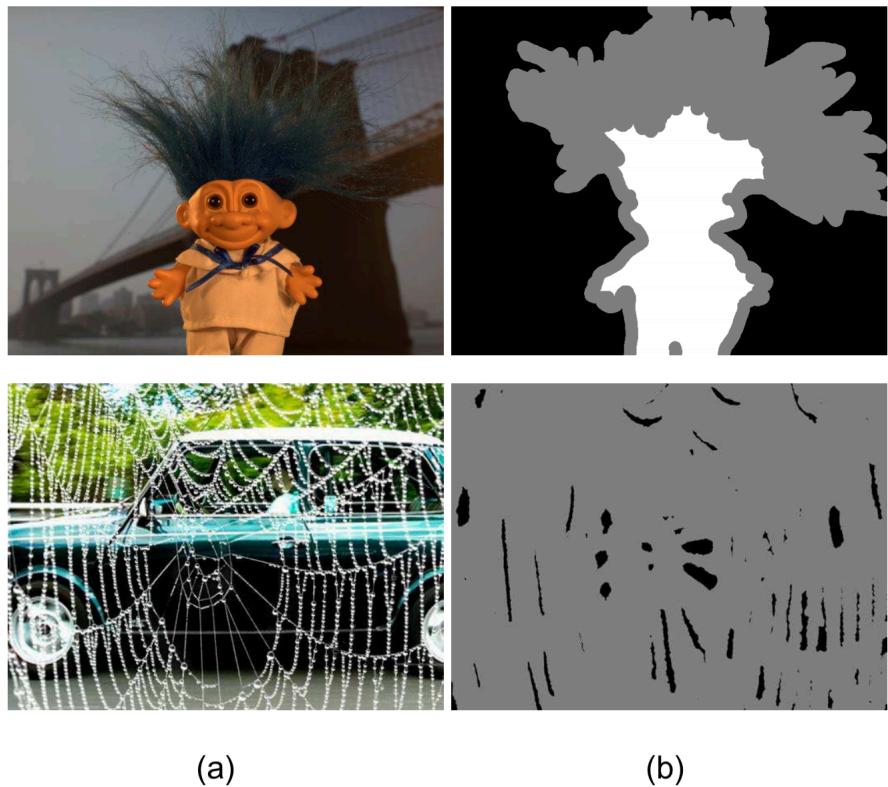


Рисунок 1 – Пример изображения с трехканальной картой [16], где (а) – изображение, (б) – трехканальная карта

позволит подавать объект (человека) без фона, что существенно упростит задачу для генеративно-состязательной сети.

Для сегментации и маттинга человека существует множество решений. К примеру, в работе [19] осуществлена сегментация и отдельно для каждого экземпляра данных нужно выделять точки скелета человека, в то время как в работе [17] используется маттинг для изображения, а значит, конечные результаты будут лучше, однако модель хорошо работает для изображений людей не в полный рост, так что для текущей задачи потребовалось бы переобучать модель на другом наборе данных с людьми в полный рост. В работе [11] для каждого входящего изображения требуется дополнительно сгенерированная трехканальная карта, а в работе [20] дополнительно требуется фотография фона от пользователя.

Как видно из перечисления, для маттирования изображения человека без зеленого экрана большинство существующих работ требуют либо вспомогательных входов, получение которых требует больших затрат, либо использо-

вания нескольких моделей, требующих больших вычислительных затрат. Следовательно, они недоступны в приложениях реального времени.

Однако были проанализированы также модели, которые не используют вспомогательных входов, с помощью квадратичной функции потерь (англ. Mean Squared Error, MSE), а также с помощью оценки среднего модуля отклонения (англ. Mean Absolute Difference, MAD), на эталонном наборе данных из 100 человеческих фотографий, собранном в статье [28]. Результаты описаны ниже в таблице 1:

Таблица 1 – Таблица существующих решений для выделения человека на изображении.

Название статьи	MSE	MAD
A Late Fusion CNN for Digital Matting [26]	0.0094	0.0158
Boosting Semantic Human Matting With Coarse Annotations [5]	0.0063	0.0114
MODNet [28]	0.0046	0.0097

По результатам была выбрана легковесная сеть MODNet [28], которая позволяет осуществлять матирование изображения человека в режиме реального времени.

1.1.2. Задача детализированной сегментации изображения человека

Детализированная сегментация изображения человека заключается в выделении на изображении человека одежды и лица. Это обусловлено задачей проекта менять позу человека, но сохранять существующие паттерны изображения, такие как одежда человека и лицо.

1.1.2.1. Наборы данных для задачи выделения одежды на изображении

Для выделения одежды на изображении человека (англ. clothes retrieval) используются наборы данных с изображениями одежды, как на человеке, так и отдельно, совместно с аннотациями к данным изображениям. Для того, чтобы проще определить положение объекта на изображении, в данном случае одежды, используются «ограничивающая рамка» (англ. bounding box), ограничивающая местоположение экземпляра одиночного объекта на картинке. Для осуществления более детальной сегментации используются маска объекта (англ. mask), то есть прямоугольная матрица принадлежности пикселя текущему объекту, а также ориентиры (англ. landmarks) — точки, определяющие контур объекта.

Для осуществления выделения одежды были проанализированы следующие наборы данных по описанным выше параметрам, подробности в таблице 2:

Таблица 2 – Таблица существующих наборов данных для выделении одежды на изображении.

Название	Кол-во изображений	Кол-во категорий	Аннот. для bboxes	Аннот. для landmarks	Аннот. для masks
DARN [9]	182K	20	7K	x	x
DeepFashion [30]	800K	50	x	120K	x
ModaNet [18]	55K	13	x	x	119K
FashionAI [24]	357K	41	x	100K	x
DeepFashion2 [27]	491K	13	801K	801K	801K

После анализа результатов, указанных в таблице 2, для обучения модели для задачи выделения одежды на изображении был выбран набор данных DeepFashion2 [27].



Рисунок 2 – Пример изображение и визуализированных аннотаций из набора данных DeepFashion2 [27]

1.1.2.2. Наборы данных для задачи выделения лица на изображении

Существует набор данных Flickr-Faces-HQ [21], который содержит 70,000 изображений высокого разрешения (1024x1024) с лицами людей разного возраста, этнической принадлежности и т.д., однако наиболее подходящим набором данных для задачи детализированной сегментации изображения является CelebAMask-HQ [2], который хоть и содержит 30,000 изображений, но для них есть аннотации для семантической сегментации частей лица на изображении, таких как «нос», «глаза», «волосы» и другие части лица.

1.2. Генеративно-состязательные сети

Генеративно-состязательные сети (англ. Generative Adversarial Networks, GANs) — алгоритм машинного обучения, входящий в семейство порождающих моделей, построенный на состязательном процессе двух нейронных сетей: генератор G , который строит приближение распределения данных, и дискриминатор D , оценивающий вероятность того, что выборка была получена из тренировочных данных, а не сгенерирована моделью G . Обучение для модели G заключается в максимизации вероятности ошибки дискриминатора D .

В генеративно-состязательных сетях обе сети, генератор и дискриминатор, обучаются за счет попытки оптимизировать целевую функцию или функцию потерь. Во время тренировки дискриминатора D мы стремимся максимизировать вероятность правильной идентификации объектов из тренировочной и сгенерированной выборок. И в то же время генератор G тренируется так, чтобы научиться генерировать наиболее реалистичные объекты.

В качестве состязательной функции потерь используется минимакс функция потерь (англ. Minimax Loss), заданная с помощью следующей формулы:

$$\mathbb{E}_x[\log(D(x))] + \mathbb{E}_z \log(1 - D(G(z))),$$

- $D(x)$ — оценка дискриминатора вероятности того, что реальный объект является реальным объектом;
- \mathbb{E}_x — математическое ожидание реальных объектов;
- $G(z)$ — сгенерированный объект с некоторым входящим шумом z ;
- \mathbb{E}_x — математическое ожидание сгенерированных объектов.

1.2.1. Условные генеративно-состязательные сети

Существует модификация генеративно-состязательных сетей – условные генеративно-состязательные сети. Условные генеративно-состязательные сети (англ. Conditional Generative Adversarial Nets, CGAN) [12] – это модифицированная версия алгоритма GAN, которая может быть сконструирована при помощи передачи дополнительных данных y , являющихся условием для генератора и дискриминатора. y может быть любой дополнительной информацией, например, меткой класса, изображением или данными из других моделей, что может позволить контролировать процесс генерации данных. Архитектура представлена на рисунке 3.

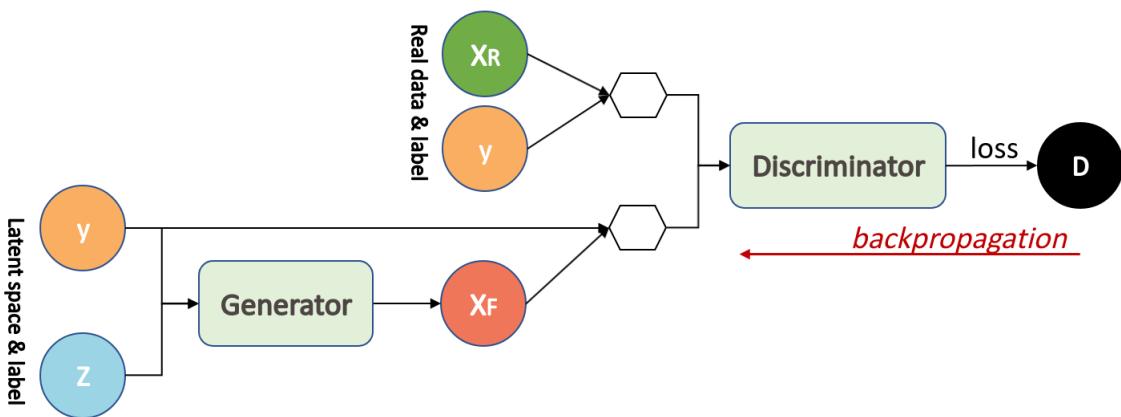


Рисунок 3 – Архитектура условных генеративно-состязательных сетей

1.2.2. Генеративно-состязательные сети с метрикой Вассерштейна для состязательной функции потерь

Существует также модификация классической состязательной функции потерь, а именно используется метрика Вассерштейна (англ. Wasserstein loss) [23], которая позволяет стабилизировать процесс обучения с использованием градиентных спусков. Функция потерь f в метрике Вассерштейна должна быть k -Липшицевой функцией, а значит, удовлетворять следующему свойству: $\forall x_1, x_2 \in R, \exists K \geq 0 : |f(x_1) - f(x_2)| \leq K * |x_1 - x_2|$.

Таким образом, состязательная функция потерь с использованием метрики Вассерштейна выглядит следующим образом: $\mathbb{E}_x[D(x)] - \mathbb{E}_z[D(G(z))]$, где D — k -Липшицевая функция. Теперь дискриминатор больше не является прямым критиком отличия поддельных образцов от настоящих. По мере того, как функция потерь уменьшается при обучении, расстояние Вассерштейна

становится меньше, и выходные данные модели генератора становятся ближе к реальному распределению данных.

1.3. Модели для генерации изображения по текстовому описанию

Создание реалистичных изображений по текстовому описанию — сложная задача в области компьютерного зрения, которая достаточно популярна и имеет множество практических применений, но современные системы искусственного интеллекта все еще далеки от этой цели.

Практически все существующие генеративно-состязательные сети для преобразования текста в изображение используют в качестве основы многослойную архитектуру, однако даже это не позволяет создавать качественные человеческие изображения на основе подписей. Генеративно-состязательные сети, такие как OPGAN [22], DFGAN [14], способны генерировать реалистично выглядящие изображения лишь для некоторых категорий, что показано на рисунке 4. Кроме того, количественная оценка этих моделей преобразования текста в изображение является сложной задачей, поскольку большинство показателей оценки оценивают только качество изображения, но не соответствие между изображением и его подписью.

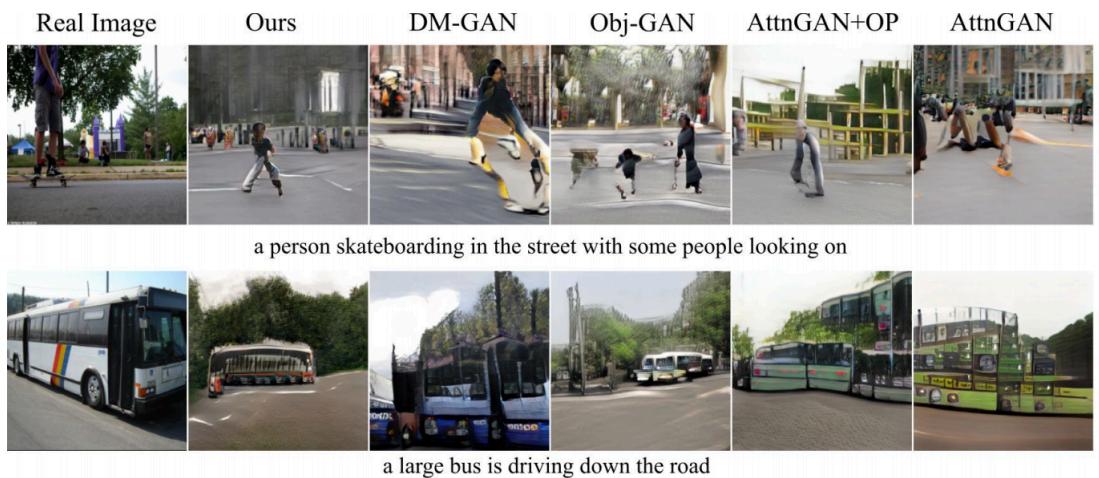


Рисунок 4 – Пример сгенерированных по тексту изображений [22]

Чтобы решить эти проблемы, в данной работе предлагается более простой, но эффективный алгоритм, который решает следующие подзадачи:

- Генерация набора точек скелета человека (позы) по текстовому описанию;
- Генерация нового изображения человека с измененной позой.

1.4. Наборы данных для генерации позы по текстовому описанию

Для генерации позы по текстовому описанию требуется набор данных, состоящий из аннотаций для точек скелета человека, а также текстовых аннотаций, описывающих, какое действие происходит.

Набор данных COCO [10], который содержит аннотации для самых разных задач, вполне покрывает выше указанные требования. Однако в текстовых аннотациях содержат размытое описание деятельности человека, что создает шум для модели.

В то время как, набор данных MPII Human Pose Dataset [15], который значительно продвинулся с точки зрения разнообразия и сложности человеческих изображений, охватывает более широкий спектр видов деятельности человека, чем в предыдущих наборах данных, включая различные виды деятельности в сфере отдыха, работы и домашнего хозяйства. Также данные изображения разбиты не только на общие категории видов деятельности, но и имеют метку подкатегории, обозначающую более точно род деятельности. К данным существуют содержательные аннотации, которые позволяют определять ограничивающие рамки (англ. bounding box) для объекта человека, а также точки скелета (англ. keypoints) человека.

Распределение категорий для данного набора данных по числу элементов описано ниже на рисунке 5.



Рисунок 5 – График распределения MPII Human Pose Dataset [15] категорий

1.5. Модели для генерации изображения человека с измененной позой

Важной задачей компьютерного зрения является также задача трансляции изображения, цель которой состоит в том, чтобы научиться строить соответствия между входным и выходным изображениями, используя тренировочные данные.

В данной работе этот подход используется для генерации нового изображения человека с измененной позой. Существует множество решений для задачи генерации человека с новой позой, для которых есть разные метрики оценки полученного результата.

Одной из важных метрик для оценки сгенерированного изображения является начальная оценка (англ. Inception Score, IS), которая отвечает на вопросы, похоже ли изображение на какой-то объект и создается ли широкий спектр объектов. Важной метрикой для оценки полученного изображения является оценка детекции объекта (англ. Detection Score, DS), которая определяет, можно ли обнаружить на изображении человека.

Краткий обзор существующих решений для генерации нового изображения с измененной позой по указанным метрикам представлен в следующей таблице 3:

Таблица 3 – Таблица существующих решений для генерации человека с измененной позой

Название модели	IS	DS
PG [8]	3.202	0.943
DPIG [7]	3.323	0.969
PATN [29]	3.209	0.973
ADGAN [25]	3.364	0.984

По результатам из таблицы было решено использовать модель, описанную в статье [25].

Выводы по главе 1

В данной главе была рассмотрена архитектура генеративно-состязательных сетей, а также их модификаций, которые будут использованы далее, а также были разобраны понятия сегментации и матирования изображений. Были проанализированы модели для сегментации изображения человека, модели для генерации человека с измененной позой и модели для

генерации изображения по текстовому описанию. Последние генерируют нереалистичные изображения плохого качества, так что теперь задача генерации изображения по текстовому описанию разбита на два шага: генерация позы по текстовому описанию и генерация изображения с измененной позой.

ГЛАВА 2. ОБЗОР ПРЕДЛОЖЕННОГО РЕШЕНИЯ ЗАДАЧИ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЯ ЧЕЛОВЕКА ПО ТЕКСТОВОМУ ОПИСАНИЮ

В этой главе речь пойдет об алгоритме для решения поставленной задачи исследования с описанием методов, с использованием теоретических и технических выкладок, позволяющих углубиться в описанный алгоритм. В решении будут использованы модели и наборы данных, которые описаны в предыдущей главе.

2.1. Описание решения задачи генерации изображения человека по текстовому описанию

Алгоритм работы был разбит на следующие подзадачи:

- а) Реализация детализированной сегментации изображения человека с использованием модели MODNet [28] для матирования изображения человека и набора данных DeepFashion2 [27] для выделения одежды на изображении;
- б) Разработка алгоритма генерации новой позы человека, релевантной текстовому описанию, с использованием MPII Human Pose Dataset [15];
- в) Реализация генерации измененного изображения по полученной позе человека с использованием ADGAN [25].

Более подробно про алгоритм генерации новой позы человека по текстовому описанию:

На вход алгоритму подается текст, который содержит описание позы человека. Текст предобрабатывается с использованием стандартных приемов: лемматизации, удаления стоп-слов, которые не несут информативной нагрузки. Для каждой категории выделяется набор ключевых слов, соответствующих ей. По тексту определяется наиболее релевантная категория, по числу вхождений ключевых слов.

Далее категория, закодированная с помощью унитарного кодирования, и сгенерированный вектор шума подается на вход генератору генеративно-состязательной сети, который возвращает набор точек скелета человека, то есть позу.

2.2. Описание решения для задачи детализированной сегментации изображения человека

В первую очередь, в данной работе нужно было осуществить сегментацию изображения человека, чтобы используемые данные далее не создавали шума для модели через внешний фон вне объекта человека.

Для осуществления сегментации изображения человека было принято решение использовать существующие модели. После анализа, описанного в таблице 1, была выбрана, как наиболее удобная и лучшая по показателям, модель MODNet [28].

Более того, учитывая то, что цель проекта состоит в том, чтобы научиться генерировать новое изображение с сохранением одежды и личности человека, то есть не изменяя внешние параметры, такие как лицо, нужно было добавить детализированную сегментацию к уже сегментированному изображению человека.

Для выделения одежды на изображении человека были проанализированы наборы данных с одеждой в таблице 2, среди которых был выбран DeepFashion2 [27], как набор данных с наибольшим количеством качественных аннотаций к изображениям с одеждой. В качестве модели для задачи выделения одежды на изображении была выбрана Mask R-CNN [6], которая является эффективной моделью в задачах сегментации экземпляров, детекции объектов и определения поз людей на фотографии (англ. *human pose estimation*). В Mask R-CNN к традиционным для алгоритмов семейства R-CNN метке класса и координатам ограничивающей рамки добавляется также маска объекта, аннотации которых есть в наборе данных DeepFashion2 [27].

Для выделения лица человека на изображения было придуман следующий алгоритм:

- а) Сначала на изображении выделяется ограничивающая рамка (англ. *bounding box*), содержащая в себе лицо человека;
- б) Далее используется алгоритм для сегментации изображения лица.

Для выделения ограничивающих рамок была использована предобученная модель [4], которая показывает хорошие результаты на фотографиях разного качества. Для сегментации лица была использована модель с архитектурой BiSeNet [1], позволяющая высококачественно осуществлять семантическую сегментацию, которая была обучена на наборе данных CelebAMask-HQ

[2], представляющий собой крупномасштабный набор данных с изображениями лиц людей с аннотациями их детализированных масок.

2.3. Описание решения для задачи генерация позы по текстовому описанию

В связи с тем, что модели для генерации изображения по текстовому описанию генерируют нереалистичные изображения, задача была разбита на два этапа, первым из которых является генерация позы по текстовому описанию. Ранее эта задача не решалась, поэтому подходящих моделей нет.

2.3.1. Набор данных для генерации позы по текстовому описанию

Для решения данной задачи был выбран набор данных MPII Human Pose Dataset [15], который представляет собой наборы точек скелета человека, подобранные по категориям.

Учитывая разброс видов деятельности, чтобы сформировать категории с наиболее похожими друг на друга позами, было осуществлено разбиение на более однозначные подкатегории, с помощью которых были сформированы следующие категории, распределение которых показано ниже на рисунке 6.

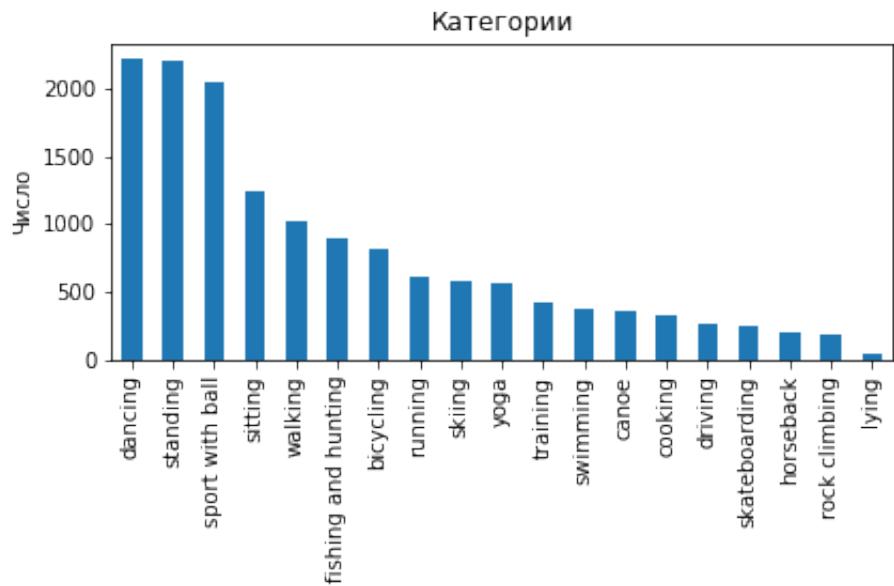


Рисунок 6 – График распределения полученных категорий

Для увеличения стабильности процесса тренировки модели, из набора данных были удалены данные с неполным количеством точек скелета. Общий размер набора данных после этого составил 10261.

Сами точки скелета представляют собой вектора $(16, 2)$, со значениями, принадлежащих диапазону $[0; 256]$. Размер изображения — $(256, 256)$. В связи с тем, что модели легче генерировать значения в диапазоне от $[0; 1]$, точки были приведены к данному формату.

Также ко всем векторам точек было применено аффинное преобразование, то есть такое преобразование, что является непрерывным, взаимно однозначным и образом любой прямой является прямая, которое учитывало центр объекта и отношение размера объекта к размеру изображения, эти данные были получены с использованием аннотаций ограниченных рамок для изображений. Это позволило нормализовать набор скелетов, представить их без поворота и увеличение или уменьшения.

2.3.2. Модель для генерации позы по текстовому описанию

После создания набора данных с точками скелета для разных категорий, задача свелась к тому, чтобы научиться по категории получать сгенерированную позу человека, релевантную заданной категории. Для этого было принято использовать архитектуру генеративно-состязательные сети.

Сформулируем входные данные для данной задачи: изначально на вход алгоритму подается категория, так называемая метка класса, на выход генерируются скелет человека из заданного категорией распределения скелетов. Учитывая то, что в алгоритм передается метка класса, в качестве архитектуры для модели было принято решение использовать архитектуру условных генеративно-состязательных сетей.

В данной алгоритме на вход генеративно-состязательной сети подается метка категории класса, закодированная с помощью унитарного кодирования. Унитарное кодирование (англ. One-Hot Encoding) – метод кодирования, при котором составляется разреженная бинарная матрица, с рангом равным количеству признаков, где единица стоит в столбце, соответствующему численному значению признака. Это позволяет избежать ситуации, когда после кодирования признаков модель может запутаться, ложно предположив, что данные связаны порядком или иерархией, которого на самом деле нет.

Однако у данной модели появились следующие проблемы:

- a) Проблема стабильности обучения (англ. non-convergence), то есть параметры модели дестабилизировались и перестали сходиться;

б) Исчезающий градиента (англ. vanishing gradient), который приводит к следующим проблемам:

- Если дискриминатор обучается плохо, то генератор не имеет обратной связи и функция потерь не может отражать реальность;
- Если дискриминатор отлично справляется с поставленной задачей, то градиент функции потерь падает почти до нуля, и обучение становится очень медленным или даже затрудненным.

Для решения этих проблем было принято использовать для состязательной функции потерь метрику Вассерштейна.

2.3.3. Вспомогательные функции потерь модели для генерации позы по текстовому описанию

В качестве дополнительного штрафа для генератора были введены вспомогательные функции потерь.

Данные на текущий момент представляют вектор точек скелета размера (16, 2). Если соединить определенные точки, получатся отрезки, являющиеся костями. Таких костей будет 16. Среди отношения длин этих отрезков можно выявить некоторые закономерности, к примеру, отношение длины локтя к отношению длины верхней части руки в среднем по набору данных около 0.00536. Данная функция потерь высчитывает вектор средних по набору данных значений всех отношений длин костей друг к другу и штрафует за отклонение от этого вектора. То есть предполагается минимизировать следующую функцию: $f(x) = \text{sum}(x_{\text{median}} - x_{\text{current}})$, где x_{median} – вектор медиан всех отношений костей, x_{current} – вектор отношений костей у текущего скелета.

Также на большинстве изображений точки, принадлежащие линии плеч или линии таза, по координате y между собой не сильно различаются. Это легко вычислить, посмотрев на среднее значение разности y -координат данных точек по датасету. Данная функция потерь высчитывает вектор средних по датасету значений разности y -координат данных точек вдоль линии плеч или вдоль линии таза и штрафует точки такого типа за отклонение по y координате.

Для разнообразия генерируемых объектов также была введена функция потерь, которая считает разницу между координатами для соседних скелетов в сгенерированном наборе и дальше минимизируется следующая функция:

$\frac{1}{sum(x-x')}$, где x – вектор сгенерированных скелетов, а x' – вектор сгенерированных скелетов с циклическим сдвигом элементов на 1.

2.3.4. Описание генератора позы по текстовому описанию

Базовая структура генератора выглядела так: шум и метка класса были сконкатенированы в самом начале сети. Для улучшение результатов, шум и метку класса было решено использовать более обособлено друг от друга, так что при входе в сеть они попадают в головы – блоки, состоящие из двух-трех линейных слоев и функции активации, после этого их выходы конкатенируются и подаются в хвост – блок, состоящий из линейных слоев и функции активации. Описанная архитектура отображена на рисунке 7.

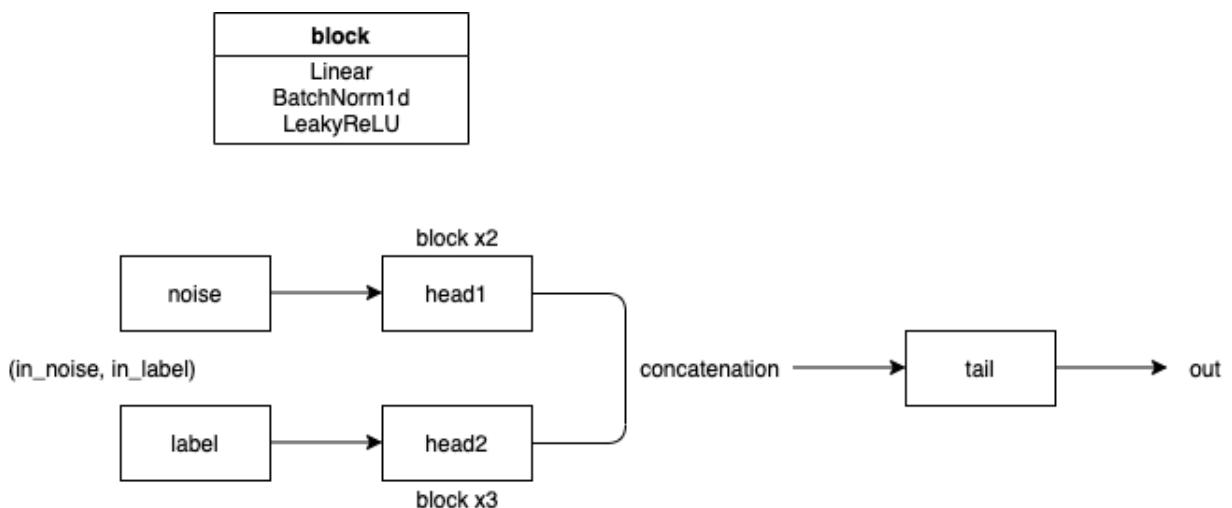


Рисунок 7 – Архитектура генератора позы по текстовому описанию

В хвосте генератора реализуются пропускаемые соединения (англ. skip connections), то есть соединения, которые пропускают какой-либо слой в нейронной сети и передают выходные данные одного уровня в качестве входных для следующих уровней (а не только для следующего). Используя пропускаемое соединение, мы обеспечиваем альтернативный путь для градиента (с обратным распространением). Описанная часть архитектуры отображена на рисунке 8.

2.3.5. Описание дискриминатора модели для генерации позы по текстовому описанию

Изначально на вход генератору подавался вектор точек скелета, однако такое решение давало плохие результаты. Поэтому для улучшения результатов

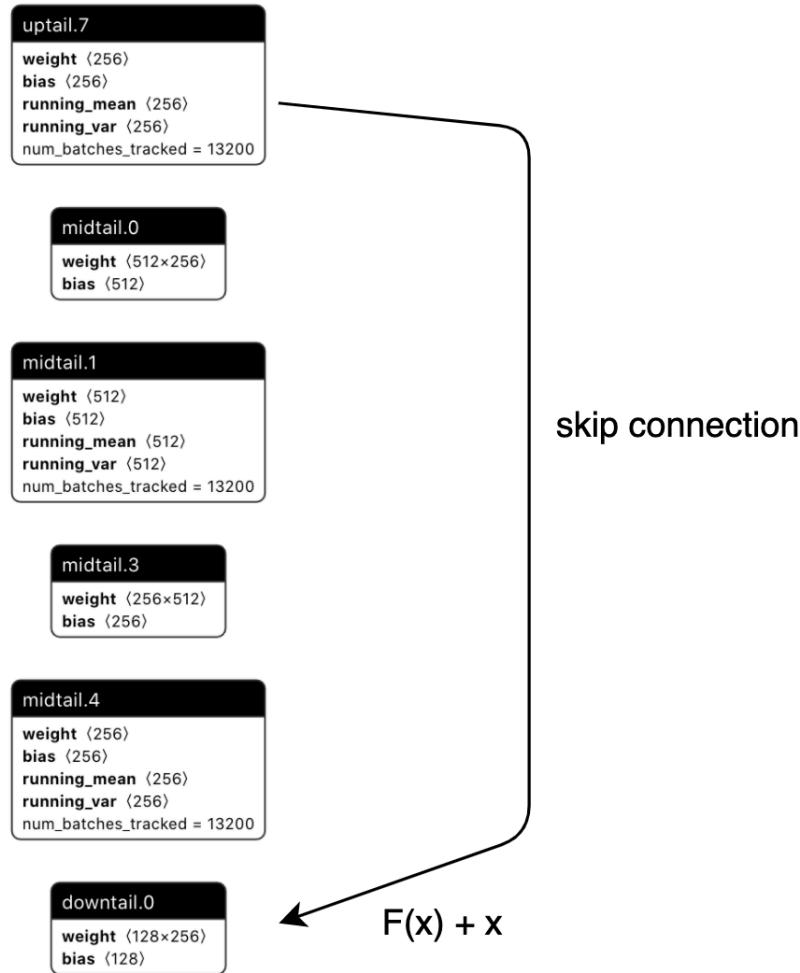


Рисунок 8 – Реализация пропускаемых соединений в архитектуре генератора

было решено рисовать гауссиану для каждой точки, то есть на вход дискриминатору подается N картинок, где N – это количество точек, на каждой из которых для соответствующей точки нарисована гауссиана в том месте, которое обозначено координатами.

Таким образом, для точки (x, y) мы получаем картинку размером $(64; 64)$, где нарисована двумерная гауссиана с центром $(x_c; y_c)$ и заданной дисперсией. Для вектора точек $(16, 2)$ мы получаем тензор размера $(16, 64, 64)$. Теперь дискриминатору подается не просто вектора точек, а картинки, по которым можно находить закономерности между точками в пространстве.

Для обработки этих изображений были добавлены в архитектуру дискриминатора сверточные слои, которые были нормализованы с помощью спектральной нормы для стабилизации обучения дискриминатора.

2.4. Описание решения для задачи генерации изображения с новой позой

Для осуществления алгоритма генерации изображения с новой позой была использована генеративно-состязательная сеть ADGAN (Attribute-Decomposed GAN) [25] для управляемой трансляции изображений, которая может создавать реалистичные изображения людей с желаемыми человеческими атрибутами (например, поза, голова, верхняя одежда), предусмотренные на различных входах источника. Архитектура описываемой модели представлена ниже на рисунке 9.

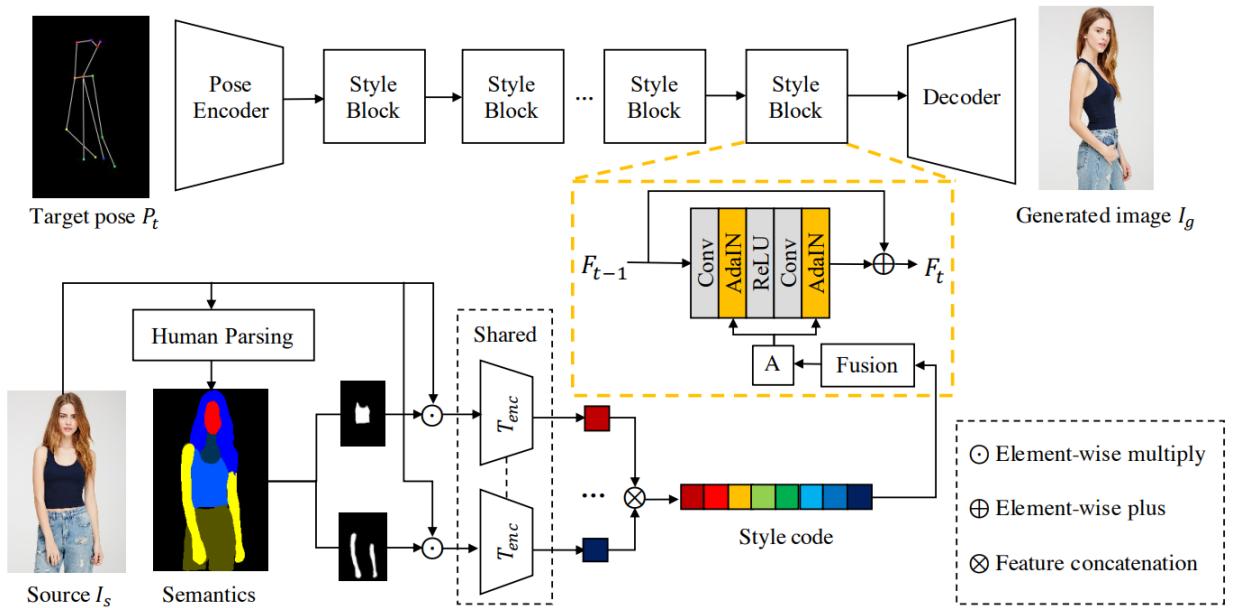


Рисунок 9 – Архитектура ADGAN[25]

Заданная поза и исходное изображение человека встроены в скрытое пространство через два независимых пути, называемых кодированием позы и кодированием разложенных компонентов, соответственно. В последнем используется ранее полученная детализированная сегментация изображения человека для разделения атрибутов компонентов и их кодирования с помощью глобального кодировщика текстур. Серия блоков стилей, оснащенных модулем слияния, вводятся для внедрения стиля текстуры исходного человека в код позы путем управления параметрами аффинного преобразования в слоях AdaIN. Наконец, желаемое изображение восстанавливается с помощью декодера.

2.5. Набор данных для алгоритма генерации позы по текстовому описанию

Для качественного обучения и тестирования модели генерации изображения человека с измененной позой, очень важно, чтобы все фотографии были высокого разрешения и на них находился ровно один человек в полный рост.

Для получения изображений были использованы различные фотостоки, для которых, были написаны программы, получающие фотографии по описанию или ключевым словам. Другим источником изображений людей в полный рост стал Fashionpedia [13], который содержал более 54 тысяч фотографий с частичным описанием. В итоге было получено более 100 тысяч изображений.

Так как мы хотим генерировать изображение человека в полный рост, то фон будет создавать шум, поэтому первым преобразованием является сегментации изображения человека. Чтобы качественно отделить человека от остального изображения, было принято использовать модель MODNet [28] на нашем наборе данных. Так как фотографии были взяты с фотостоков и готового набора данных Fashionpedia [13], то текстовое описание и ключевые слова также были получены из соответствующих источников. Пример изображения можно увидеть ниже на рисунке 10.

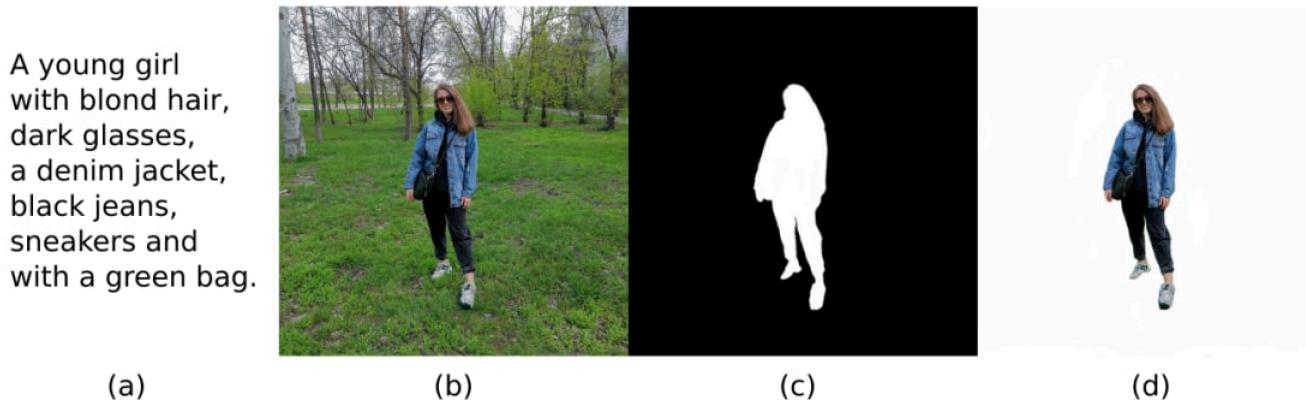


Рисунок 10 – Экземпляр из собранного набора данных, где (а) – текстовая аннотация, (б) – реальное изображение, (с) – альфа-маска, (д) – сегментированное изображение

Выводы по главе 2

Подводя итог, была реализована модель для генерации позы по текстовому описанию с использованием архитектуры условной генеративно-состязательной сети с заданными вспомогательными функциями. Был собран

тестовый набор данных с использованием моделей для удаления внешнего фона вне человека. Были встроены в решения существующие модели для детализированной сегментации и генерации изображения человека с новой позой.

ГЛАВА 3. ОБЗОР ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ ДЛЯ ЗАДАЧИ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЯ ПО ТЕКСТОВОМУ ОПИСАНИЮ

В этой главе пойдет речь о результатах, полученных в процессе реализации решения, и о тестировании предложенного решения.

3.1. Полученные результаты для задачи детализированной сегментации изображения человека

Для задачи сегментации изображения человека была успешно применена модель MODNet [28], в результате чего был получен тестовый набор данных изображений человека без внешнего фона для тестирования.

Примеры изображений представлены ниже на рисунке 11:



Рисунок 11 – Примеры сегментированных человеческих изображений

Для выделения одежды на изображении была обучена Mask-RCNN [6] на наборе данных DeepFashion2 [27].

3.2. Полученные результаты для задачи генерации позы человека по текстовому описанию

Была реализована условная генеративно-состязательная сеть, которая была обучена на MPII Human Pose Dataset [15], с помощью которой по описаным выше категориям генерируется реалистичная поза (точки скелета).

По графикам можно отследить, что в процессе обучения дискриминатор начинает предсказывать лучше для реальных объектов, чем для сгенерированных, что можно увидеть на рисунке 12.

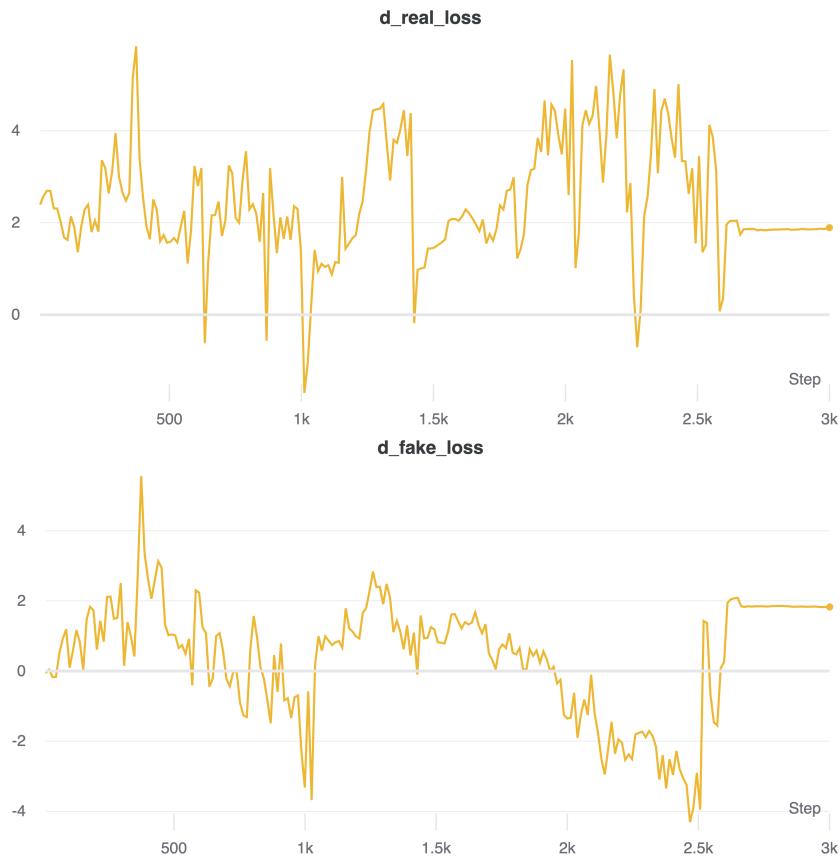


Рисунок 12 – График функций потерь для дискриминатора

Примеры сгенерированных поз можно увидеть на рисунке 13.

3.3. Тестирование результатов для задачи генерации позы человека по текстовому описанию

Для тестирования результатов был проведен опрос, где ассессорам нужно было определить, подходит ли предложенное изображение текстовому описанию. Изображение выбиралось из собранного набора данных, как самое близкое к сгенерированной позе по текстовому описанию. На рисунке 14 можно увидеть пример текстового описание с релевантной ему сгенерированной позой и наиболее близким изображением из собранного набора данных.

Результаты опроса показали, что с вероятностью 0,79 с погрешностью 0,1 опрашиваемым нравится подобранное изображение по сгенерированной позе, релевантной текстовому описанию.

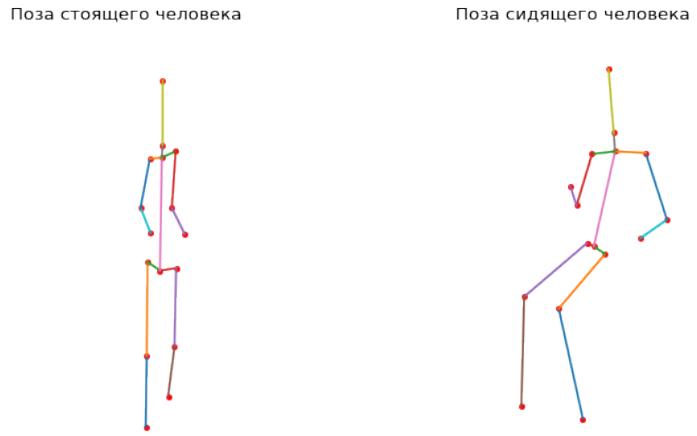


Рисунок 13 – Примеры сгенерированных поз



Рисунок 14 – Пример генерации позы по текстовому описанию

Выводы по главе 3

В данной главе были описаны результаты проделанной работы, которые оказались положительными. Решение было протестировано с помощью собранного набора данных. Результаты проведенного опроса показали, что изображения с позой, близкой к сгенерированной, соответствуют текстовому описанию.

ЗАКЛЮЧЕНИЕ

В данной работе был произведен обзор существующих решений для задачи генерации изображения человека по текстовому описанию, однако удовлетворяющих результатов не было найдено. Был предложен и реализован алгоритм для генерации изображения человека с измененной позой по текстовому описанию, результаты которого получились удовлетворительными, за счет использования декомпозиции в подходе к решению данной задачи.

Для достижения поставленной цели была реализована модель для генерации позы человека по текстовому описанию, а также были внедрены модели для детализированной сегментации изображения человека и генерации изображения человека с новой позой. Для обучения моделей и тестирования был собран набор данных высокого разрешения, что положительно повлияло на результаты.

Разработанная модель для генерации позы человека по текстовому описанию расширяет возможности в разных сферах, позволяя с помощью текста обозначить желаемую для изображения позу и получить ее на изображении.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 *Changqian Yu C. G.* BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation [Электронный ресурс]. — 2020. — URL: <https://arxiv.org/abs/2004.02147>.
- 2 *Cheng-Han Lee Z. L.* MaskGAN: Towards Diverse and Interactive Facial Image Manipulation [Электронный ресурс]. — 2019. — URL: <https://arxiv.org/abs/1907.11922>.
- 3 *Ian J. Goodfellow J. P.-A.* Generative Adversarial Networks [Электронный ресурс]. — 2014. — URL: <https://arxiv.org/abs/1406.2661>.
- 4 *Jian Li Y. W.* DSFD: Dual Shot Face Detector [Электронный ресурс]. — 2019. — URL: <https://arxiv.org/abs/1810.10220>.
- 5 *Jinlin Liu Y. Y.* Boosting Semantic Human Matting With Coarse Annotations [Электронный ресурс]. — 2020. — URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Liu_Boosting_Semantic_Human_Matting_With_Coarse_Annotations_CVPR_2020_paper.html.
- 6 *Kaiming He G. G.* Mask R-CNN [Электронный ресурс]. — 2017. — URL: <https://arxiv.org/abs/1703.06870>.
- 7 *Liqian Ma Q. S.* Disentangled Person Image Generation [Электронный ресурс]. — 2018. — URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Ma_Disentangled_Person_Image_CVPR_2018_paper.html.
- 8 *Liqian Ma X. J.* Pose Guided Person Image Generation [Электронный ресурс]. — 2017. — URL: <https://arxiv.org/abs/1705.09368>.
- 9 *M. Hadi Kiapour X. H.* Where to Buy It: Matching Street Clothing Photos in Online Shops [Электронный ресурс]. — 2015. — URL: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Kiapour_Where_to_Buy_ICCV_2015_paper.html.
- 10 *Maire T.-Y. L.* Microsoft COCO: Common Objects in Context [Электронный ресурс]. — 2014. — URL: https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

- 11 *Marco Forte F. P. F, B.* Alpha Matting [Электронный ресурс]. — 2020. — URL: <https://arxiv.org/abs/2003.07711>.
- 12 *Mehdi Mirza S. O.* Conditional Generative Adversarial Nets [Электронный ресурс]. — 2014. — URL: <https://arxiv.org/abs/1411.1784>.
- 13 *Menglin Jia M. S.* Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset [Электронный ресурс]. — 2020. — URL: https://link.springer.com/chapter/10.1007/978-3-030-58452-8_19.
- 14 *Ming Tao H. T.* DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis [Электронный ресурс]. — 2020. — URL: <https://arxiv.org/abs/2008.05865v2>.
- 15 *Mykhaylo Andriluka L. P.* 2D Human Pose Estimation: New Benchmark and State of the Art Analysis [Электронный ресурс]. — 2014. — URL: https://openaccess.thecvf.com/content_cvpr_2014/html/Andriluka_2D_Human_Pose_2014_CVPR_paper.html.
- 16 *Ning Xu B. P.* Deep Image Matting [Электронный ресурс]. — 2017. — URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Xu_Deep_Image_Matting_CVPR_2017_paper.html.
- 17 *Quan Chen T. G.* Semantic Human Matting [Электронный ресурс]. — 2018. — URL: <https://arxiv.org/abs/1809.01354v2>.
- 18 *Shuai Zheng F. Y.* ModaNet: A Large-Scale Street Fashion Dataset with Polygon Annotations [Электронный ресурс]. — 2018. — URL: <https://arxiv.org/abs/1807.01394>.
- 19 *Song-Hai Zhang R. L.* Pose2Seg: Detection Free Human Instance Segmentation [Электронный ресурс]. — 2019. — URL: <https://arxiv.org/abs/1803.10683>.
- 20 *Soumyadip Sengupta V. J.* Background Matting: The World is Your Green Screen [Электронный ресурс]. — 2020. — URL: <https://arxiv.org/abs/2004.00626v2>.
- 21 *Tero Karras S. L.* A Style-Based Generator Architecture for Generative Adversarial Networks [Электронный ресурс]. — 2018. — URL: <https://arxiv.org/abs/1812.04948>.

- 22 *Tobias Hinz Stefan Heinrich S. W.* Semantic Object Accuracy for Generative Text-to-Image Synthesis [Электронный ресурс]. — 2015. — URL: <https://arxiv.org/abs/1910.13321v2>.
- 23 *Weng L.* From GAN to WGAN [Электронный ресурс]. — 2019. — URL: <https://arxiv.org/abs/1904.08994>.
- 24 *Xingxing Zou X. K.* FashionAI: A Hierarchical Dataset for Fashion Understanding [Электронный ресурс]. — 2019. — URL: https://openaccess.thecvf.com/content_CVPRW_2019/html/FFSS-USAD/Zou_FashionAI_A_Hierarchical_Dataset_for_Fashion_Understanding_CVPRW_2019_paper.html.
- 25 *Yifang Men Y. M.* Controllable Person Image Synthesis with Attribute-Decomposed GAN [Электронный ресурс]. — 2020. — URL: <https://arxiv.org/abs/2003.12267>.
- 26 *Yunke Zhang L. G.* A Late Fusion CNN for Digital Matting [Электронный ресурс]. — 2019. — URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_A_Late_Fusion_CNN_for_Digital_Matting_CVPR_2019_paper.html.
- 27 *Yuying Ge R. Z.* DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images [Электронный ресурс]. — 2019. — URL: <https://arxiv.org/abs/1901.07973>.
- 28 *Zhanghan Ke K. L.* Is a Green Screen Really Necessary for Real-Time Portrait Matting? [Электронный ресурс]. — 2020. — URL: <https://arxiv.org/abs/2011.11961>.
- 29 *Zhen Zhu T. H.* Progressive Pose Attention Transfer for Person Image Generation [Электронный ресурс]. — 2019. — URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_Progressive_Pose_Attention_Transfer_for_Person_Image_Generation_CVPR_2019_paper.html.

- 30 *Ziwei Liu P. L.* DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations [Электронный ресурс]. — 2016. — URL: https://openaccess.thecvf.com/content_cvpr_2016/html/Liu_DeepFashion_Powering_Robust_CVPR_2016_paper.html.