Санкт-Петербургский государственный университет

Михайлов Станислав Александрович

Выпускная квалификационная работа

Система распознавания эмоций и аномалий в выражениях лица человека и сгенерированных лиц с помощью методов машинного обучения

Уровень образования: Бакалавриат

Направление 02.03.01 Математика и компьютерные науки

Основная образовательная программа СВ.5152.2019 Математика, алгоритмы и анализ данных

Научный руководитель:
Доцент,
Факультет математики и
компьютерных наук СПбГУ
Авдюшенко Александр Юрьевич

Рецензент:
Доцент,
Федеральное государственное
автономное образовательное
учреждение высшего образования
«Санкт-Петербургский политехнический
университет Петра Великого»,
Перезябов Олег Аркадьевич

Санкт-Петербург
2023

Оглавление

Введение **4**

1. Обзор предметной области **7**

1.1. Базовые понятия предметной области 7

1.2. Постановка задачи распознавания лицевых выражений человека 7

1.3. Обзор датасетов для распознавания лицевых выражений 7

1.4. Модели для распознавания лицевых выражений 9

1.4.1. Классические алгоритмы компьютерного зрения 9

1.4.2. Нейросетевые модели 11

1.4.3. EfficientNet v1/v2 13

1.4.4. DAN 15

1.4.5. Другие нейросетевые методы 17

1.5. Модели для генерации изображений 17

1.5.1. Модели, основанные на архитектуре генеративно-состязательных нейросетей 18

1.5.2. DualStyleGan 21

1.6. Задача генерации аватаров 23

1.6.1. Применение генеративных моделей для создания аватаров 23

1.7. Детекция аномалий 24

1.7.1. Архитектура "Сиамская модель" 24

1.8. Мотивация работы. Формулировка цели и постановка задач 26

2. Разработка моделей распознавания эмоций и генерации аватаров **28**

2.1. Выбор датасета для задачи распознавания лицевых выражений людей . 28

2.2. Детали обучения моделей для распознавания лиц людей 28

2.2.1. Предобработка данных 28

2.2.2. EfficientNet 29

2.2.3. DAN 31

2.3. Генерация аватаров с помощью генеративных моделей 34

2.3.1. Выбор целевого стиля для генерации аватаров 34

2.3.2. Технические детали генерации датасета аватаров 35

2.3.3. Примеры несоответствия эмоций исходного изображения и аватара 36

3. Система детектирования аномалий **37**

3.1. Подготовка данных и моделей для реализации системы детектирования аномалий 37
3.1.1. Разметка датасета аватаров 37
3.1.2. Дообучение модели распознавания эмоций на изображениях аватаров 38
3.2. Разработка системы детектирования аномалий 38
3.2.1. Первая версия пайплайна 39
3.2.2. Вторая версия пайплайна 41

Список литературы **42**

Приложение **48**

Введение

За последние десятилетия технологические достижения оставили глубокий след на нашем образе жизни. Эти достижения все больше проникают в различные сферы жизни человека. В последнее время одним из особенно актуальных направлений является внедрение машинного обучения и искусственного интеллекта (ИИ) в человеческие области, которые раньше казались недостижимыми для применения компьютерных алгоритмов.

С развитием проектов, связанных с метавселенными - виртуальными пространствами, где все участники наделены аватарами, - искусственный интеллект становится неотъемлемым инструментом для создания гармоничного и плодотворного общения между людьми и виртуальными сущностями. Особенно можно отметить ряд задач, связанных с моделированием внешности человека. Последние годы множество компаний и исследовательских лабораторий работают над созданием метавселенных - виртуальных миров, в которых люди будут иметь возможность взаимодействовать друг с другом с помощью гарнитур виртуальной реальности (VR) и дополненной реальности (AR)[1]. Внутри метавселенной каждому человеку соответствует аватар - 2D или 3D графическое (цифровое) представление пользователя[2]. Поэтому, высокий интерес вызывает сфера распознавания эмоций лица человека(FER) на основе глубокого анализа изображений и видео, так как одной из самых важных частей человеческого взаимодействие - это общение. При неправильной трансляции эмоции или выражения человеческого лица его аватара, слова человека могут быть восприняты неправильно.

Объем и количество крупных баз данных по задаче распознавания эмоций значительно расширились за последние два десятилетия RAF-DB [39], Affectnet [2], что привело к значительному улучшению точности распознавания некоторых моделей сверточных нейронных сетей (CNN). Однако, несмотря на недавние выдающиеся результаты, FER до сих пор считается сложной задачей из-за нескольких причин:

- **Глобальные факторы**. Существующие методы FER не до конца распознают глобальные факторы входных изображений из-за ограничения сверточных локальных рецептивных полей;
- **Межклассовое сходство**. Несколько категорий выражений часто включают похожие изображения с небольшими различиями между ними;
- **Внутриклассовое неравенство**. Изображения из той же категории выраже-

[1] https://wikipedia.org/wiki/Metaverse
[2] https://wikipedia.org/wiki/Avatar_(computing)

ния лица могут существенно отличаться друг от друга, например, цвет лица, пол, фон изображения и возраст человека различается в зависимости от экземпляра;

- **Чувствительность моделей**. Различия в качестве и разрешении изображения могут часто ставить под угрозу эффективность сетей глубокого обучения при использовании без необходимых мер предосторожности. Изображения из наборов данных, которые похожи на изображения из реальной жизни(in the wild) и других наборов данных FER представлены в широком диапазоне с разными размерами изображений. Следовательно, для FER важно обеспечить стабильную производительность в разных масштабах.

Графическое представление человека можно получить с помощью инструментов для рендеринга или нарисовать вручную. Однако это требует навыков и времени, между тем современные методы глубокого обучения позволяют получать высококачественные изображения, уменьшая затраты по времени и также не имея особых требований к навыкам пользователя.

Задача генерации графического представления людей является весьма сложной задачей, так как достаточно тяжело точно передавать детали лица и его выражение, ведь человеческое лицо представляет собой целую систему из множества мимических мышц. В данных момент нет готовых метавселенных, в которых можно было бы сгенерировать готовые аватары по лицу человека и собрать хороший датасет для обучения и улучшения моделей, которые работают с цифровыми представлениями лица человека. Несмотря на это, в настоящее время существуют генеративные нейросетевые модели, которые достаточно успешно справляются с image-to-image генерацией качественных изображений. Такие модели в том числе используются для создания аватаров по исходному изображению человека. Однако они имеют свои достоинства и недостатки в зависимости от используемой архитектуры: разнообразие генерируемых данных, частота дискретизации, скорость генерации и т.д. Поэтому, многие из них могут недостаточно точно передавать детали выражения эмоций. Однако, такие данные могут помочь построить модель, которую в будущем можно будет активно применить в реальных системах.

Целью данной работы является построение системы, которая будет распознавать несоответствия(аномалии) между выражениями лица человека и лица его сгенерированного аватара используя современные методы машинного обучения.

Структура работы

В главе 1 будет описана постановка задачи распознавания лицевых выражений, постановка задачи генерации аватаров для обучения системы и будет рассказано о пайплайне детекции аномалий между эмоциями реальных и сгенерированных лиц. Также, приводится обзор предметной области, датасетов и существующих решений для подобных задач, их сравнение.

В главе 2 будут описаны детали реализации, особенности и применения моделей FER. Будут приведены использующиеся датасеты, гиперпараметры моделей, метрики во время обучения и валидации, сравнение разных версий моделей. Также, будет описана модель для генерации датасета аватаров, необходимого для построения системы детекции аномалий, с примерами сгенерированных лиц.

В главе 3 будут описаны детали реализации системы детекции аномалий, приведены результаты и анализ распознавания.

1. Обзор предметной области

1.1. Базовые понятия предметной области

Ниже приведены определения терминов, которые в дальнейшем будут использоваться в работе.

Аватар — графическое представление пользователя.

Пайплайн детекции аномалий —- система типа конвейер, которая на вход получает пару изображений реального лица человека и аватара, а на выходе выдает ответ, присутствует ли аномалия.

FER —- задача распознавания лицевых эмоций

1.2. Постановка задачи распознавания лицевых выражений человека

Распознавание выражений лица (FER) — это задача компьютерного зрения, направленная на выявление и классификацию эмоциональных выражений, изображенных на человеческом лице. Цель состоит в том, чтобы автоматизировать процесс определения эмоций в режиме реального времени, анализируя различные черты лица, такие как брови, глаза, рот и другие черты, и сопоставляя их с набором эмоций, таких как:

- Удивление
- Страх
- Отвращение
- Счастье
- Грусть
- Гнев

Зачастую к вышеупомянутым шести эмоциям прибавляют седьмую - нейтральное выражение.

1.3. Обзор датасетов для распознавания лицевых выражений

Для обучения нейронных сетей под задачу распознавания эмоций лица требуется досточно большое количество размеченных данных.

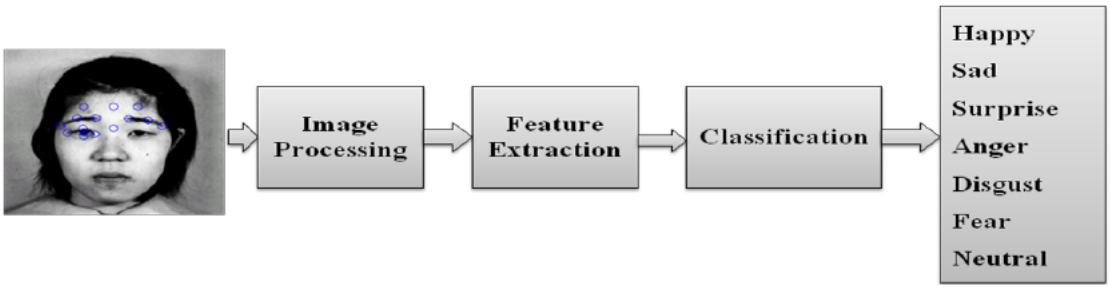


Рис. 1: Задача распознавания лицевых выражений

В работе [38] приведен список публичных датасетов, доступных для применения при решении задачи распознавания лицевых выражений. Наиболее часто используемые датасеты и сравнение их характеристик приведены в Таблице 1.

Наиболее крупный датасет EmotioNet [8] был получен с помощью разметки 25 000 изображений людьми, после чего остальные изображения были классифицированы с помощью автоматических алгоритмов разметки. Датасет AffectNet [2] был размечен вручную и на данный момент является одним из крунейших датасетов для распознавания эмоций, размеченных людьми. Этот датасет содержит разбиение на 8 эмоций и часто используется в качестве бенчмарка, но является закрытым и представляется только по запросу.

Отдельно стоит поговорить про некоторые общие для FER датасетов трудности. Смещение данных и неконсистентная аннотация весьма распространены среди многих датасетов для FER в силу разнообразия условий сборки данных и субъективности разметки. В связи с этим нередко модели имеют достаточно невысокую обобщающую способность, и при оценке на других датасетах с новыми данными они показывают существенно более низкое качество распознавания эмоций людей. Еще одной очень распространненой чертой FER датасетов является их несбалансированность. Сильные негативные эмоции, как злость или страх, проявляются значительно реже, чем счастье, к примеру. К тому же при разметке такие сложные эмоции иногда можно определить неверно. Для решения проблемы несбалансированности распределения классов лицевых выражений можно использовать аугментацию данных или синтез новых изображений на стадии предобработки датасета. Другим вариантом может служить использование различных функций потерь, чувствительных к количеству примеров каждого класса (например весовые коэффициенты, которые зависят от доли примеров класса по отношению ко всему датасету).

Датасет	Тип	Объем	Источник	Распределение эмоций
CK+	Видео	593 видеофайла	Лабораторная сборка	6 базовых эмоций, нейтральное выражение и сложные эмоции
MMI	Видео и изображения	740 изображений и 2 900 видеофайлов	Лабораторная сборка	6 базовых эмоций и нейтральное выражение
JAFFE	Изображения	213 изображений	Лабораторная сборка	6 базовых эмоций и нейтральное выражение
TFD	Изображения	112 234 изображений	Лабораторная сборка	6 базовых эмоций и нейтральное выражение
FER-2013	Изображения	35 887 изображений	Интернет	6 базовых эмоций и нейтральное выражение
AFEW 7.0	Видео	1 809 видеофайлов	Фильмы	6 базовых эмоций и нейтральное выражение
SFEW 2.0	Изображения	1 766 изображений	Фильмы	6 базовых эмоций и нейтральное выражение
BU-3DFE	Изображения	2 500 изображений	Лабораторная сборка	6 базовых эмоций и нейтральное выражение
Oulu-CASIA	Последовательности изображений	2 880 изображений	Лабораторная сборка	6 базовых эмоций
RaFD	Изображения	1 608 изображений	Лабораторная сборка	6 базовых эмоций, нейтральное выражение и презрение
KDEF	Изображения	4 900 изображений	Лабораторная сборка	6 базовых эмоций и нейтральное выражение
EmotioNet	Изображения	1 000 000 изображений	Интернет	23 эмоции
RAF-DB	Изображения	29 672 изображений	Интернет	6 базовых эмоций, нейтральное выражение и 12 сложных эмоций
AffectNet	Изображения	450 000 изображений	Интернет	6 базовых эмоций и нейтральное выражение
ExpW	Изображения	91 793 изображений	Интернет	6 базовых эмоций и нейтральное выражение

Таблица 1: Сравнительная таблица публичных датасетов для задачи FER [38]

1.4. Модели для распознавания лицевых выражений

Существует достаточно много алгоритмов для распознавания лицевых выражений. До активного использования нейросетевых подходов применялись классические алгоритмы компьютерного зрения. Такие методы подразумевают выделение признаков из изображений или кадров видео с помощью различных способов. В работе [46] приводится обзор классических и нейросетевых методов для решения задачи FER, а также их сравнение.

1.4.1. Классические алгоритмы компьютерного зрения

Классические алгоритмы компьютерного зрения, как правило, состоят из нескольких этапов: предобработка, извлечение признаков из полученного изображения и применение классификатора либо ансамбля классификаторов.

Предобработка помогает снизить влияние таких факторов, как внутриклассовые различия и межклассовые сходства лицевых выражений, небольшие изменения внешнего вида лица, позы головы, угла съемки, освещенности и контраста. Для начала необходимо детектировать (локализовать) лицо на изображении, для этого могут ис-

пользоваться метод классические подходы: метод Виолы-Джонса [56], гистограмма направленных градиентов (HoG) [18, 15] или же более эффективные и современные нейросетевые решения, например Single Shot multibox Detector (SSD) [47] или Multi-Task Cascaded Convolutional Networks (MTCNN) [32]. Далее по предсказанным точкам прямоугольника (bounding box) детектора выполняется обрезание и масштабирование изображения, то есть приведение к единому разрешению. После чего производится выравнивание лица относительно опорных изображений, что позволяет уменьшить различия внутри классов эмоций. Дополнительным этапом предобработки может быть регулировка контрастности, что помогает сглаживать изображения, избавляться от шума и улучшать насыщенность, приводя все изображения примерно к одинаковому уровню освещенности. Одними из таких методов являются выравнивание гистограммы (histogram equalization) [57] или линейное контрастное растяжение (linear contrast stretching) [42]. Выбор правильной многоэтапной предобработки позвляет ускорить работу классификатора и повысить его точность. Пример предобработки для изображения приведен на Рисунке 2.

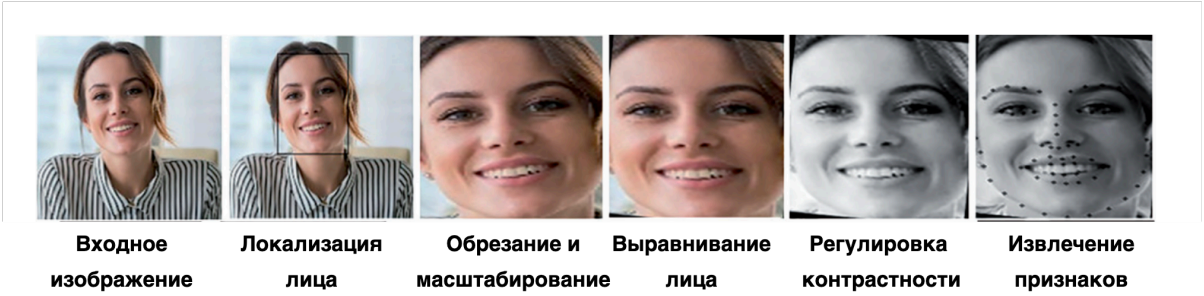


Рис. 2: Пример многоэтапной предобработки изображения для извлечения признаков с помощью классических алгоритмов [46]

Для извлечения визуальных признаков из предобработанного изображения выделяют следующие группы методов: на основе геометрических особенностей, на основе текстур и на основе глобальных и локальных объектов. Первая группа методов направлена на извлечение информации о местоположении наиболее важных геометрических объектов, например частей лица человека. Основная идея этих методов заключается в том, что расстояния между объектами и их координаты могут использоваться в качестве признаков для классификации лицевых выражений. Примерами таких дескрипторов для задачи FER являются SIFT [9], гистограмма направленных градиентов, активная модель внешнего вида (AAM) [31] или курвлет преобразование (Curvelet Transform) [20]. Следующий набор методов использует текстурные особенности изображения. К таким методам применительно к распознаванию лицевых

выражений относятся: фильтры Габора [9], локальный бинарный шаблон (LBP) [49], локальное фазовое квантование (LPQ) [19], локальный дескриптор Вебера [36]. Третья группа методов, которые основаны на глобальных и локальных объектах, включают в себя метод главных компонент (PCA) [55], линейный дискриминативный анализ (LDA) [55] или использование оптического потока (optical flow) [58]. Наиболее полезными методами являются методы на основе текстурных особенностей, однако они хуже справляются с окклюзиями по сравнению с геометрическими методами. Также существуют гибридные методы для извлечения признаков лицевых эмоций.

Заключительным этапом является классификация эмоций на основе извлеченных признаков. Разделяют традиционные подходы к классификации и искуственные нейронные сети. Наиболее используемыми при решении задачи распознавания лицевых выражений из традиционных методов являются расстояние Хаусдорфа [7], метод опорных векторов (SVM) [55, 13, 54], марковские модели (Hidden Markov Models [55]) и деревья решений [54]. Среди искусственных нейросетей в качестве классификаторов находят применение обычные многослойные сети прямого распространения [23], сверточные нейронные сети (CNN) [51, 33], реккурентные нейронные сети (RNN) [3] и многие другие модели с различными архитектурами.

1.4.2. Нейросетевые модели

В последнее десятилетие, для решения задач компьютерного зрения в большинстве случаев используют нейросетевые методы на основе сверточных нейронных сетей(CNN). Довольно подробный обзор нейросетевых решений задачи распознавания лицевых выражений приведен в работе [38].

Предобработка в случае нейросетевых моделей похожа на препроцессинг при использовании классических подходов и состоит из следующих шагов: детекция лица, обрезание и масштабирование, выравнивание, регулировка контрастности и аугментация данных (Рисунок 3). Все шаги, кроме аугментации данных могут выполняться аналогично классическим методам и с помощью различных инструментов, упомянутых ранее.

Особое внимание стоит уделить аугментации данных, так как глубокие нейронные сети требуют достаточно много обучающих данных для успешного решения задачи распознавания лицевых выражений. Однако большинство публичных датасетов не обладают достаточным количеством данных, чтобы достичь хорошей обобщающей способности моделей. Для того, чтобы увеличить количество тренировочных примеров и сделать их более вариативными, тем самым уменьшая вероятность переобучиться под небольшой датасет, и применяют аугментацию. Разделяют так называемую ауг-

Рис. 3: Препроцессинг изображения для нейросетевых классификаторов эмоций [38]

ментацию «на лету» и оффлайн аугментацию [38]. Первый вид аугментации подразумевает случайные изменения изображения во время тренировочного шага. Например вырезание (crop) участка изображения или его отражение для каждого изображения внутри батча. Оффлайн аугментация зачастую устроена более сложно, она способна существенно расширять исходный набор данных и вносить вариативность, делая модель более устойчивой к небольшим изменениям входных изображений. Наиболее частыми трансформациями являются вращение, смещение, сжатие, добавление шума различной природы, изменение контраста, яркости, цветов (color jittering), отражение. Комбинации этих и многих других операций можно использовать с помощью достаточно удобного фреймворка компьютерного зрения Albumentations [1]. Однако существуют и куда более продвинутые методы аугментации, как CutOut, CutMix [10], MixUp, MixAugment [44]. Последние два метода были использованы для задачи распознавания лицевых выражений авторами в работе [44]. Визуально этот тип модификаций обучающих примеров приведен на Рисунке 4. Техника MixUp направлена на создание «виртуальных» примеров $\tilde{x}$, с меткой $\tilde{y}$ на основании случайно выбранных реальных сэмплов x_i, x_j из обучающей выборки с метками y_i, y_j согласно уравнению 1. Параметр $\lambda \sim B(\alpha, \alpha) \in [0, 1]$ (Бета-распределение) и $\alpha \in (0, \infty)$.

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1 - \lambda) x_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda) y_j \end{aligned} \tag{1}$$

Тогда в качестве функции потерь для «виртуальных» примеров предлагается использовать категориальную кросс-энтропию (CCE).

$$L^v_{CCE} = E_{\overline{y},\tilde{y}}[-\tilde{y} \log \overline{y}] \tag{2}$$

где $\overline{y}$ - предсказанная вероятность для $\tilde{x}$.

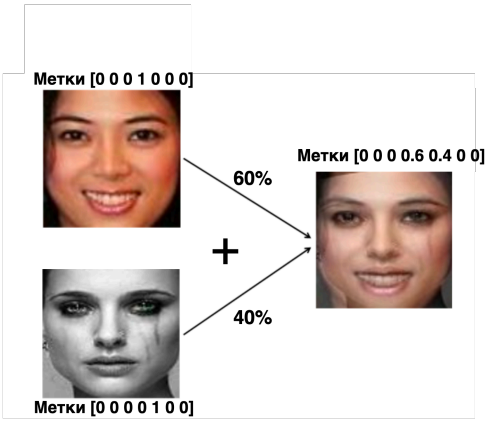


Рис. 4: Создание «виртуального» сэмпла на основе реальных с помощью аугментации MixUp

Методы, связанные со смешением изображений (Mix-методы) расширяют тренировочное распределение за счет включения априорных знаний о том, что линейные интерполяции вектора признаков должны приводить к линейной интерполяции соответствующих метток классов. Авторы [44] провели исследование влияния сложных техник аугментации во время обучения нейросетей и выяснили, что подобные методы могут применяться для повышения качества классификации эмоций в том числе.

Методы глубокого обучения направлены на извлечение наиболее релевантной высокоуровневой информации с помощью иерархических архитектур множественных нелинейных преобразований. Существует большое разнообразие архитектур нейросетей, которые могут применяться для получения признаков с помощью обучаемых параметров моделей. В работе [38] применительно к задаче FER упоминаются сверточные нейронные сети (CNN), генеративно-состязательные сети (GAN) [21], реккурентные сети (RNN) и сети с долгой краткосрочной памятью (LSTM).

1.4.3. EfficientNet v1/v2

EfficientNet — класс новых моделей, который получился из изучения масштабирования (скейлинг, scaling) моделей и балансирования между собой глубины и ширины количества каналов сети, а также разрешения изображений в сети. Авторы статьи предлагают новый метод составного масштабирования (compound scaling method), который равномерно масштабирует глубину/ширину/разрешение с фиксированными пропорциями между ними. Из существующего метода под названием «Neural Architecture Search» [16] для автоматического создания новых сетей и своего собствен-

ного метода масштабирования авторы получают новый класс моделей под названием EfficientNets [52].

Они лучше и намного экономнее предыдущих сеток. На ImageNet [30] EfficientNet-B7 достигает state-of-the-art , будучи при этом в 8.4 раза меньше и в 6.1 раз быстрее в использовании, чем лучшая на тот момент по точности ConvNet. Также, она показало хорошие результаты на других датасетах – EfficientNets получили state-of-the-art на 5 из 8 наиболее популярных датасетов.

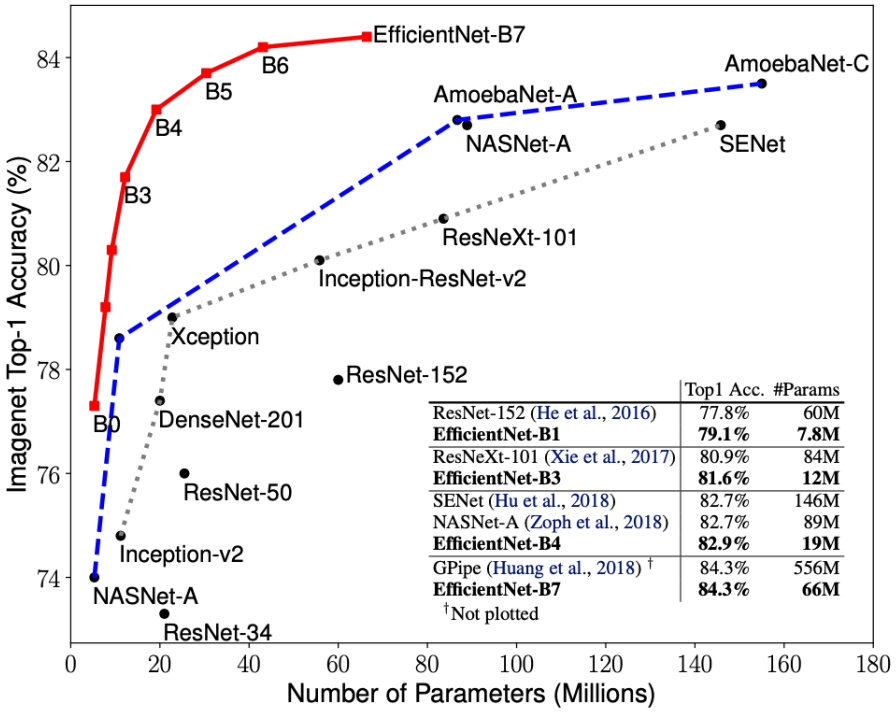
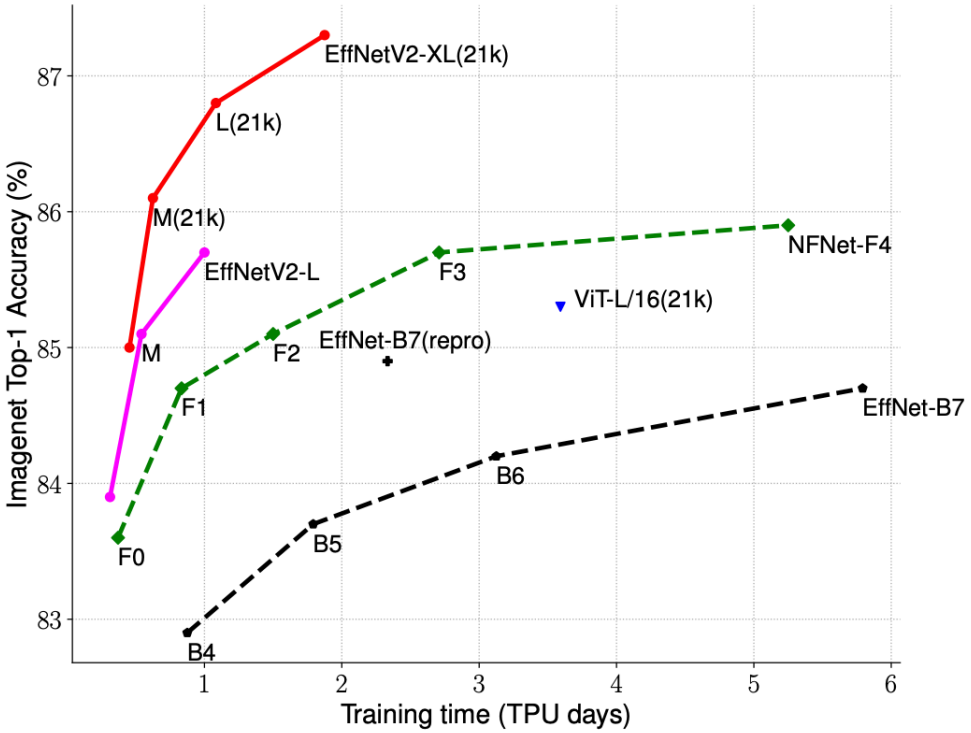


Рис. 5: График точности и количества параметров сети на датасете ImageNet [52]

EfficientNetV2 [53] делает еще один шаг вперед по сравнению с EfficientNet [52], увеличивая скорость обучения и эффективность параметров. Эта сеть создается с использованием комбинации масштабирования (ширина, глубина, разрешение) и поиска нейронной архитектуры. Основная цель — оптимизировать скорость обучения и эффективность параметров. Кроме того, на этот раз пространство поиска также включало новые сверточные блоки, такие как Fused-MBConv. В итоге авторы получили архитектуру EfficientNetV2, которая намного быстрее предыдущих и более новых современных моделей и намного меньше (до 6,8 раз). Это показано на рисунке 6.



(a) Training efficiency.

	EfficientNet (2019)	ResNet-RS (2021)	DeiT/ViT (2021)	EfficientNetV2 (ours)
Top-1 Acc.	84.3%	84.0%	83.1%	83.9%
Parameters	43M	164M	86M	24M

(b) Parameter efficiency.

Рис. 6: График точности и количества параметров сети на датасете ImageNet [52]

1.4.4. DAN

Модель DAN [14] показывает высокие результаты точности в задаче FER. Данный метод реализует несколько голов внимания (multiple attention heads) [6] и гарантирует, что они захватывают полезные аспекты выражения лица без наложения. На рисунке 7 можно увидеть, какие области покрывает модель с 1 головой и 4 головами.

Конкретно, в данной работе используются три подсети (рисунок 8), включая нейронную сеть для извлечения параметров (FCN), сеть внимания с несколькими головами (MAN) и сеть объединения внимания (AFN). Сначала модель извлекает и группирует вектор параметров, полученных из основной сети, встроенную в FCN, где применяется AffinityLoss для увеличения расстояний между классами при уменьшении расстояний внутри классов. С помощью MAN строются параметры, основанные на

одновременном внимания к нескольким областям лица, где несколько головок внимания, каждая из которых включает блок пространственного внимания и блок внимания по каналам. Наконец, векторы выходных признаков из MAN передаются в AFN для вывода оценок по классам.

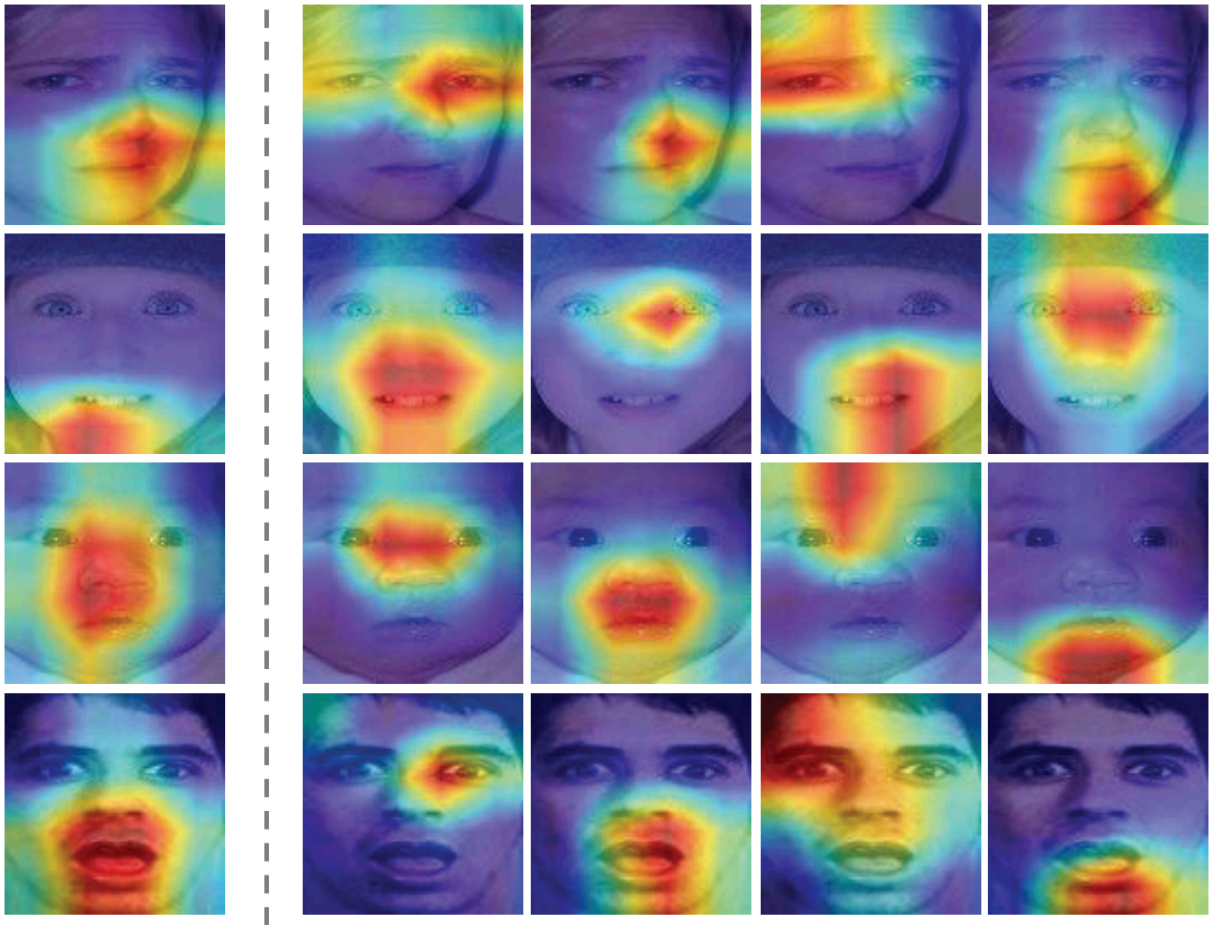


Рис. 7: Сравнение ключевых областей для модели с 1 головой внимания и 4 головами внимания с помощью Grad-CAM++ визуализации [22]

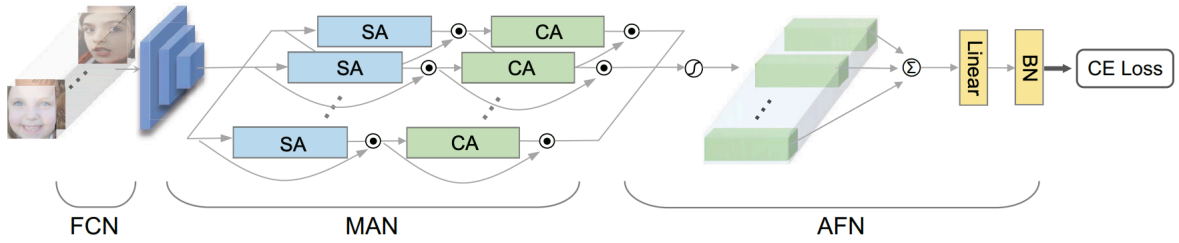


Рис. 8: Архитектура модели DAN

1.4.5. Другие нейросетевые методы

В последних работах, например [37] авторы предлагают применять относительно новую архитектуру - визуальные трансформеры (ViT) [29], которым даже удавалось показать лучшее качество распознавания эмоций по сравнению с CNN на некоторых датасетах.Модели с улучшенной архитектурой визуальных трансформеров Swin Transformer[50] показали более точные метрики классификации на бенчмарк датасете ImageNet1k по сравнению с ViT и составили конкуренцию EfficentNet[52] и RegNet[45].

Большинство сверточных моделей, которые показывают наилучшие результаты в решении задачи классификации лицевых выражений, используют в качестве backbone нейросети предобученные вариации моделей с архитектурой ResNet [11]. Такие модели обучаются на огромных бенчмарк датасетах, например ImageNet [30] с более чем 14 миллионами изображений или датасет Microsoft-Celeb-1M [40] с около 10 миллионами изображений лиц людей, для дальнейшего применения на более узких задачах (downstream tasks) в похожей области - классификация эмоций, возраста, пола и т.д. Такие state-of-the-art модели для задачи FER на датасетах лиц людей, как EAC[35] или DAN[14], используют механизм внимания в слоях после backbone нейросети. Также нередко для достижения наиболее точных результатов используется сложные аугментации, наподобие вышеупомянутых MixUp и MixAugment.

Также стоит подвести итог, что при достаточном наборе обучающих данных глубокие нейронные сети показывают себя лучше классических подходов и способны обнаруживать эмоции с более высокой скоростью и точностью [46].

1.5. Модели для генерации изображений

В общем смысле генеративный подход в машинном обучении, в отличие от дискриминативного, моделирует не условную вероятность $p(y|x)$, а совместную - $p(x, y)$, где x - наблюдаемые данные, а y - таргеты. То есть генеративные модели решают в некотором смысле более общую задачу по сравнению с дискриминативными.

Область генерации изображений является важной частью компьютерного зрения и компьютерной графики. С помощью наиболее современных моделей стало возможным создание фотореалистичных изображений с высоким разрешением со способностью осуществлять контроль над семантическими атрибутами изображений.

Постановка задачи генерации изображений заключается в создании модели машинного обучения, которая способна создавать новые изображения, соответствующие определенным критериям. В общем случае дана вероятностная модель генерации изображений $p_{\text{model}}(\mathbf{x})$, где $\mathbf{x}$ - это изображение. Задача состоит в том, чтобы обучить

эту модель на основе обучающего набора изображений $\mathcal{X} = \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$, так чтобы она генерировала изображения, которые похожи на изображения из обучающего набора.

В работе [25] достаточно хорошо описаны принципы генерации изображений и общие идеи устройства генеративных моделей. Так, генеративная модель G, параметризованная θ, принимает на вход случайный шум $\mathbf{z}$ и выдает сэмпл $G(\mathbf{z}, \theta)$. Таким образом, выход модели может быть интерпретирован как пример из распределения $G(\mathbf{z}, \theta) \sim p_g$. В тренировочной выборке x имеются данные из распределения p_{data} и цель генеративной модели G во время обучения - аппроксимировать p_{data} с помощью p_g.

Современные модели позволяют достаточно просто и быстро генерировать качественные изображения с заданными пользователем параметрами (стиль, фон, объекты на изображении и т.д.). На Рисунке 9 изображены основные архитектуры моделей, которые используются для генерации изображений: GAN, VAE и диффузионные модели.

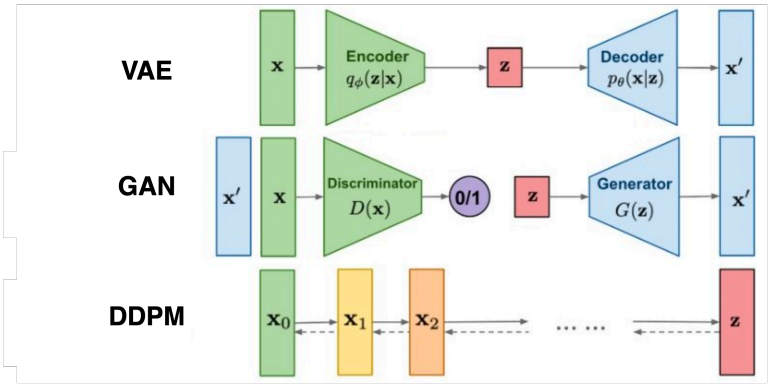


Рис. 9: Архитектуры моделей для генерации изображений

1.5.1. Модели, основанные на архитектуре генеративно-состязательных нейросетей

Генеративно-состязательные нейросети (GAN) впервые были описаны в работе [21]. Они также состоят из двух компонент - генератор и дискриминатор, которые идейно «соперничают» друг с другом: задача генератора получить изображение, инициализированное случайным шумом, такое, чтобы дискриминатор не смог его отличить от изображения из референсного датасета, а дискриминатор в свою очередь обучается классифицировать настоящие и синтетические данные как можно более

точно. Общая схема архитектуры генеративно-состязательных нейросетей изображена на Рисунке 10. Генератор обычно представляет собой нейросеть с устройством автокодировщика.

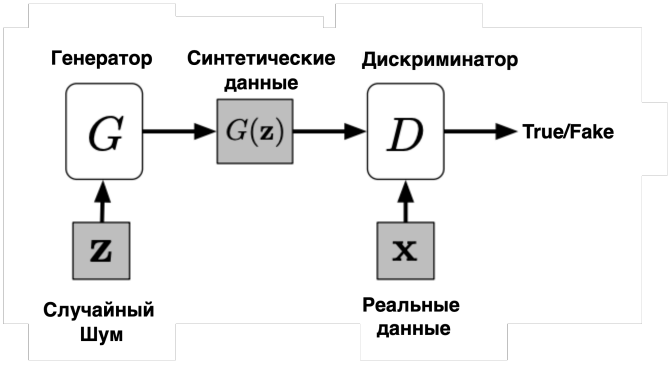


Рис. 10: Общая схема архитектуры GAN

Дискриминатор D и генератор G обучаются попеременно, в то время как веса другого фиксированы. Далее опишем общий алгоритм обучения моделей с GAN-подобной архитектурой.

Шаг обучения дискриминатора, в то время как веса генератора заморожены:

1. Выбирается батч эталонных изображений x
2. Генерируется батч изображений $x' = G(z)$ для случайных $z \sim N(0, 1)$
3. Дискриминатор предсказывает $p = D(x)$ и $p' = D(x')$
4. Происходит обратное распространение ошибки кросс-энтропии весов D, метка для $p' = 0$

Шаг обучения генератора, в то время как веса дискриминатора заморожены:

1. Генерируется батч изображений $x' = G(z)$ для случайных $z \sim \mathcal{N}(0, 1)$
2. Дискриминатор предсказывает $p' = D(x')$
3. Происходит обратное распространение ошибки кросс-энтропии весов G, метка для $p' = 1$

Такой способ обучения отличается от привычного и часто называется минимакс оптимизацией

$$\min_G \max_D V(D, G) = E_x[\log D(x)] + E_z[\log(1 - D(G(z)))] \tag{3}$$

Генератор сфокусирован на минимизации слагаемого $\log(1-D(G(z)))$, а дискриминатор на максимизации $\log D(x)$. При таком подходе "оптимальным" является решение, при котором распределение $G(z)$ идеально повторяет истинное распределение x и, соответственно, дискриминатор выдает одинаковые вероятности $D(x) = D(x') = \frac{1}{2}$.

У классического подхода GAN есть несколько недостатков: затухание градиента (diminished или vanishing gradient), схлопывание мод распределений (mode collapse) и сходимость обучения. Под исчезновением градиента понимается ситуация, когда дискриминатор справляется со своей задачей слишком хорошо, и градиент генератора затухает, и он практически не обучается. Проблемы сходимости связаны с тем, что многие гарантии и теоремы, доказанные для обычных нейросетевых моделей, которые обучаются с помощью градиентного спуска, не выполняются для моделей типа GAN. Поэтому задача оптимизации не будет поиском локального или глобального минимума, а скорее точки равновесия. Проблема схлопывания мод распределений заключается в том, что генератор не учится на самом деле воспроизводить исходное распределение, а учится обыгрывать дискриминатор, поэтому у него ограниченное количество выходов и изображения могут часто повторяться. Эта проблема является одной из основных и с ней можно успешно бороться, например, для этого в Wasserstein GAN (WGAN) [5] используется метрика Вассерштайна.

Несмотря на описанные выше проблемы при обучении GAN, существует много техник и модификаций, благодаря которым эта архитектура (хоть и с существенными изменениями) и по сей день очень часто используется и в некоторых задачах остается state-of-the-art решением. Применительно к задаче генерации изображений людей и человекоподобных аватаров наиболее выделяются модели, описанные далее. Модель StyleGAN [34] внесла существенный вклад в развитие задачи image synthesis и стала одной из лучших на момент выхода применительно к безусловной генерации изображений, в том числе лиц людей. Её улучшенная версия StyleGAN2 [4] также часто упоминается в различных статьях и считается одной из самых применимых GAN моделей для unconditioned face generation.

Таким образом, генеративно-состязательные сети и их различные вариации по типу Conditional GANs (CGAN) [41] сильно превзошли по качеству и реалистичности генерируемых изображений предыдущие модели, в том числе VAE. Однако из-за трудностей во время обучения работать с такими моделями сложнее работать и в зависимости от специфики задачи нужно подбирать подходы для борьбы с недостатками GAN.

1.5.2. DualStyleGan

DualStyleGAN [43] представляет собой генеративно-состязательную сеть (GAN), основанную на архитектуре StyleGAN. Модель DualGAN разработана для решения задач трансформации изображений и переноса стилей, таких как перевод изображений из одного домена в другой.

В основе модели DualStyleGAN лежат две генеративно-состязательные сети (GAN [21]), каждая из которых состоит из генератора и дискриминатора. Обучение DualStyleGAN происходит путем оптимизации обеих сторон GAN, так что они учатся переводить изображения из одного домена в другой и обратно.

Последовательность работы DualStyleGAN выглядит следующим образом:

- Генератор G преобразует исходное изображение A (домен A) в изображение B' (домен B).
- Дискриминатор D1 сравнивает полученное изображение B' с реальными изображениями из домена B и вычисляет функцию потерь.
- Аналогично, генератор G2 преобразует исходное изображение B (домен B) в изображение A' (домен A).
- Дискриминатор D2 сравнивает полученное изображение A' с реальными изображениями из домена A и вычисляет функцию потерь.
- Обе пары генераторов и дискриминаторов обновляются на основе полученных функций потерь.

Главное отличие между DualGAN и другими аналогичными моделями, такими как CycleGAN, заключается в архитектуре генератора. Как и в случае с обычной моделью StyleGAN [34], генераторы DualStyleGAN используют технологию адаптивного стиля. Эта технология предполагает использование операций преобразования стилей для переноса стилевых признаков, обучаемых аналитически. Это позволяет модели DualStyleGAN контролировать уровень абстракции и структуры изображений, что, в свою очередь, приводит к гибкости при синтезе и генерации.

Одно из ключевых преимуществ DualStyleGAN заключается в генерации более качественных изображений с точки зрения сохранения семантики и переноса стиля, по сравнению с другими методами переноса стиля и трансформации изображений. Это достигается благодаря использованию архитектуры StyleGAN и ее усовершенствованных генераторов, которые дают возможность более точно контролировать структурные и стилевые характеристики генерируемых изображений.

В отличие от обычной модели StyleGAN, которая используется для генерации случайных изображений определённого стиля, Dual StyleGAN применяется в более широком спектре задач, таких как перенос стиля, преобразование стилей изображений, аугментация данных и прочее.

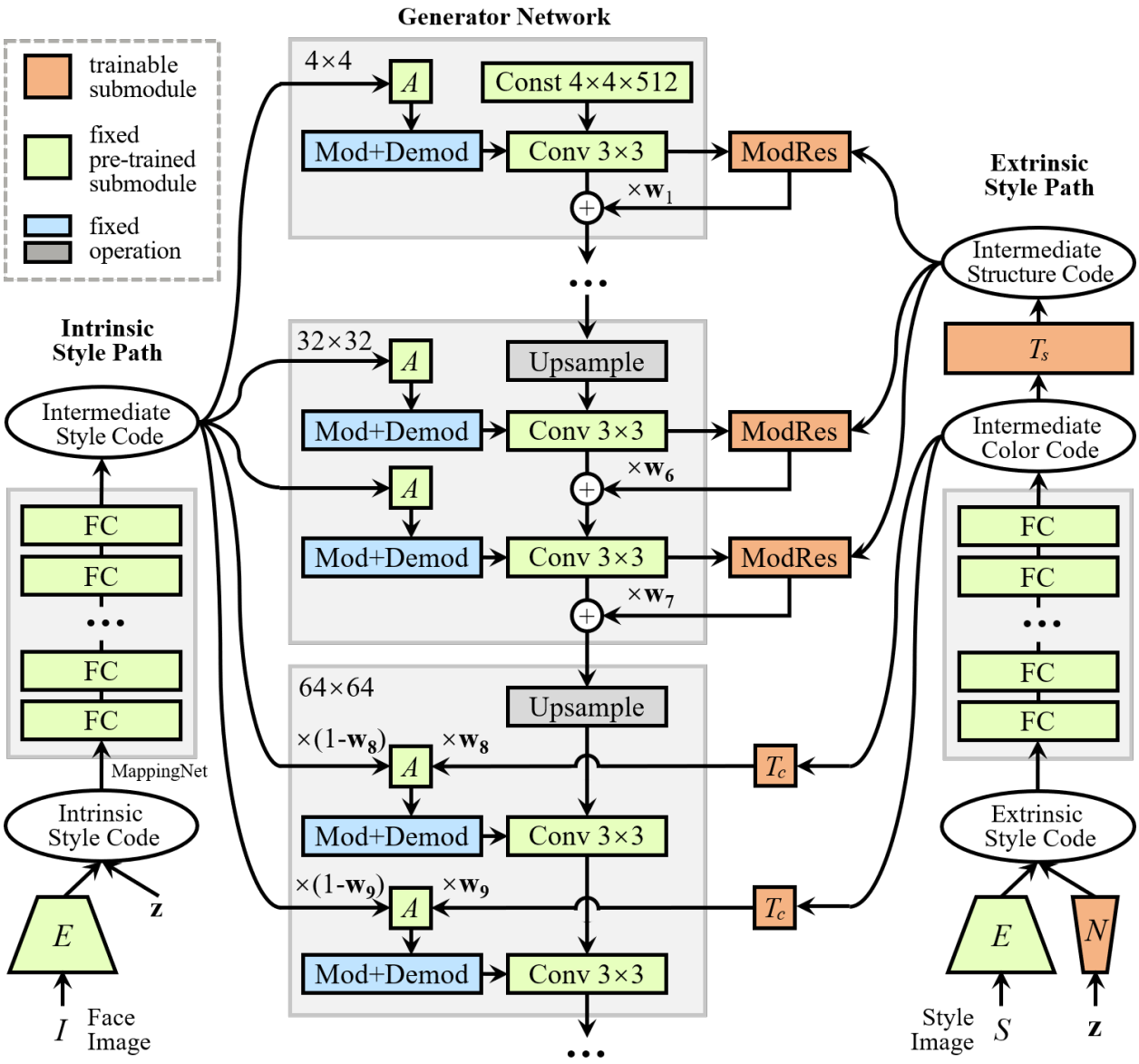


Рис. 11: Детализация архитектуры DualStyleGan [43]

На рисунке 11 показаны детали сети DualStyleGAN [43]. Внутренняя стуктура стиля и сеть генератора образуют стандартный StyleGAN [34] и остаются фиксированными во время тонкой настройки(fine-tuning). Структура внутреннего стиля принимает код внутреннего стиля единичного гауссовского шума $z \in R^{1\times512}$. Внешние коды стилей также могут быть взятием выборки через сеть дискретизации N путем отображе-

ния единичных гауссовских шумов к внешнему распространению стиля. Формально при наличии изображения лица I и художественного портретного изображения S, передача образцового стиля осуществляется с помощью $G(E(I), E(S), w)$, где $w \in R^{18}$ — вектор весов для гибкой комбинации двух стилей, по-умолчанию установлено значение 1. Генерация художественного портрета реализуется через $G(z_1, N(z_2), w)$. Когда $w = 0$, G деградирует в стандартный g для генерации лиц, то есть $G(z, \cdot, 0) \sim g(z)$.

StyleGAN [34] обеспечивает иерархическое управление стилем, где слои с высоким разрешением и низким разрешением моделируют низкоуровневый цветовой стиль и высокоуровневый стиль формы соответственно, который является вспомогательным для дизайна структуры внешнего стиля.

1.6. Задача генерации аватаров

С помощью генеративных моделей можно решать множество задачч: генерация лица (face generation), модификацая лиц (face modification), перенос стиля (style transfer) и многие другие.

1.6.1. Применение генеративных моделей для создания аватаров

Для разработки и анализа системы детектирования аномалий, нужно получить датасет аватаров.

Задачу генерации аватаров по исходному изображению человека можно рассматривать как задачу переноса стиля (image-to-image style transfer) с изображения реального человека с сохранением семантических особенностей - структуры выражения лица, поза, пропорции лица на изображении. Для решения задачи StyleTransfer хорошо подходят генеративные модели.

Сгенерированные аватары должны обладать следующими характеристиками:

- Структура лица и сохранение пропорций. Аватар должен быть похож на человека с исходного изображения, из которого он генерировался
- Сохранение выражения лица, сгенерированный аватар должен сохранять основные черты и выражение лица человека, изображенного на исходном изображении
- Сохранение позы и ракурса, чтобы добиться максимальной схожести с первоначальным изображением
- Высокое разрешение, чтобы генерируемое изображение было достаточно хорошего качества

На Рисунке 12 схематично изображен пример создания аватара из исходного изображения с помощью генеративной модели.

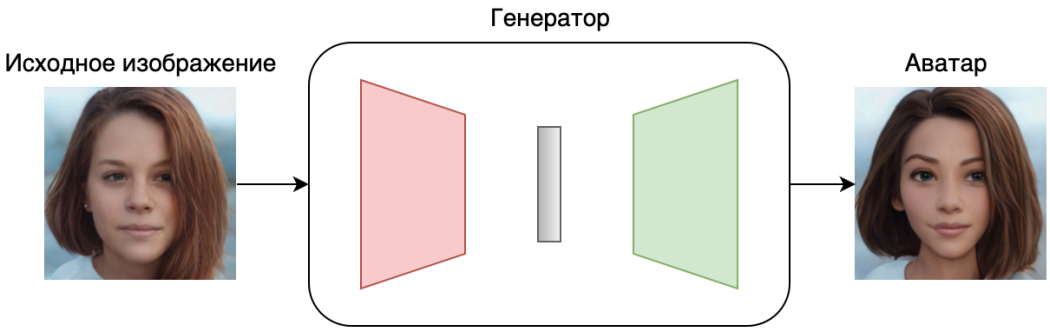


Рис. 12: Высокоуровневая схема генерации аватара из исходного изображения

С помощью хорошего генератора, можно полностью сгенерировать датасет аватаров из исходного. Данный датасет необходим, чтобы дообучить модели FER для распознавания эмоций на аватарах. Также, аватары понадобятся для разработки модели детекции аномалий из двух изображений в качестве тестовой и обучающей выборки.

1.7. Детекция аномалий

Детекция аномалий(несоответствий) между эмоциями на двух изображениях реального лица и аватара, подразумевает сравнение или анализ двух разных изображений, на которых изображены люди с разными лицевыми выражениями. Целью этого анализа является определение и оценка различий между эмоциями на этих изображениях.

Для построения такой системы с помощью моделей распознавания эмоций можно рассмотреть несколько подходов:

- Сравнение меток эмоций между исходным изображением и аватаром из выходов двух моделей FER
- Использовать вектора параметров, которые извлекают FER модели для распознавания эмоций, и сравнивать полученные вектора. Архитектура данного подхода напоминает сиамскую нейронную сеть.

1.7.1. Архитектура ”Сиамская модель”

Рассмотрим подробнее архитектуру сиамской нейронной сети.

Сиамские сети состоят из двух идентичных нейронных сетей, каждая из которых принимает одно из двух входных изображений. Затем последние слои двух сетей передаются в функцию потерь, которая вычисляет сходство между двумя векторами. Общую архитектуру сиамской нейронной сети можно увидеть на рисунке 13.

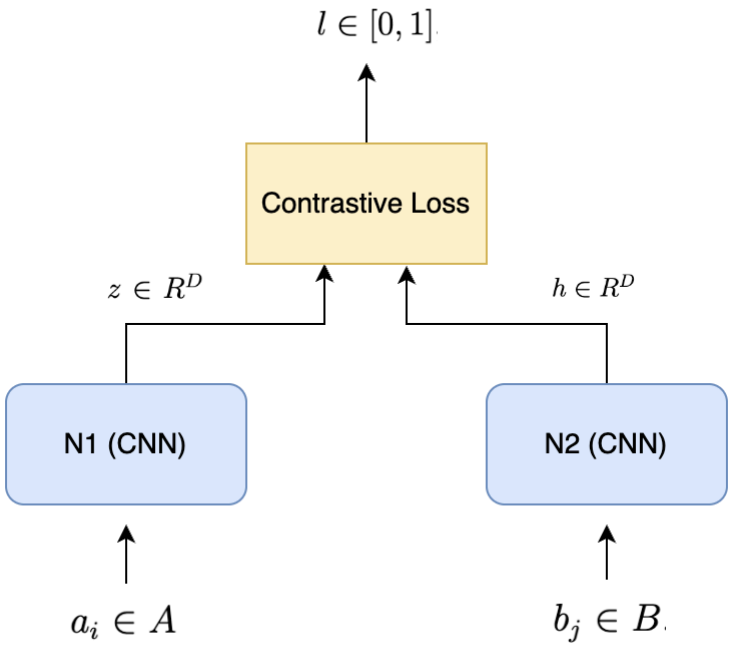


Рис. 13: Обобщенная архитектура сиамской модели

Технически, для задач обработки изображений, подсети сиамской модели могут быть сверточными нейронными сетями (CNN) или основной частью(backbone) большой предобученной модели.

Пусть имеется A и B - два набора изображений. Подсети получают на вход вектора $a_i \in A$ и $b_j \in B$. Предположим, подсети на последнем слое выдают на выходе вектор размером $D = 512$. Тогда, на выходе подсети извлекают вектора признаков $z \in R^D$, $h \in R^D$. Далее, для вычисления сходства двух векторов используется Contrastive Loss Function [24]:

$$CSLoss(a_i, b_j, y_{ij}) = (1 - y_{ij})\frac{1}{2}D^2 + y_{ij}\frac{1}{2}\{\max(0, m - D\}^2$$

где $y_{ij} \in \{0, 1\}$ - метка принадлежности одному классу, $m > 0$ - это заранее заданный предел, а D - евклидово расстояние между выходами подсетей от входов a_i, b_j:

$$D(a_i, b_j) = \sqrt{(z - h)^2}$$

где $z = N_1(a_i), h = N_2(b_j)$ - выходы подсетей сиамской модели.

На выходе, сиамская модель выдает оценку $l \in [0, 1]$, насколько различны вектора.

1.8. Мотивация работы. Формулировка цели и постановка задач

В ближайшем будущем большую роль в цифровом мире будут играть метавселенные. Огромная часть жизни в метавселенной это социальное взаимодействие людей друг с другом.

При генерации аватара человека в метавселенной нельзя быть уверенным, что он идеально передаст все детали выражения лица человека, некоторые части могут быть искажены или отображены неправильно. Это может стать большой проблемой, если при взаимодействии людей друг с другом они будут неправильно интерпретировать эмоции друг друга, из-за этого могут возникать недопонимания, конфликты или вовсе, цифровое пространство будет неудобным для людей, и они начнут покидать его.

Цель данной работы: разработать модели распознавания эмоций с использованием современных технологий в области машинного обучения. На основе FER моделей придумать и разработать систему, которая будет распознавать несоответствия(аномалии) между выражениями лица человека и лица его сгенерированного аватара.

Данная система позволит предотвратить недопонимания. Система может быть встроена в метавселенную в режиме реального времени, например, при возникновении аномалии в выражении лица человека, над ним будет отображаться информация, что эмоция передана некорректно, и выводить реальную эмоцию человека(полученную из части системы, которая распознает эмоции реального лица человека).

Кроме того, данную систему можно будет использовать как метрику качества генерации аватаров для метавселенных, чем меньше будет аномалий в передаче эмоций, тем более качественная генерация аватара в метавселенной.

Для достижения данной цели решаются следующие задачи:

1. Проанализировать современные модели обработки изображения, поставить эксперименты по их улучшению и реализовать на их основе модели для распознавания эмоций лица с использованием датасета под задачу FER

2. Проанализировать существующие решения генерации аватаров для выявления качественных моделей и выбрать наиболее подходящую

3. Применить выбранную генеративную модель для создания датасета аватаров и разметить полученный датасет

4. Дообучить модели FER на датасете аватаров и оценить качество распознавания
5. Придумать и разработать архитектуру пайплайна по детектированию аномалий на основе полученных FER моделей и проанализировать его работу

2. Разработка моделей распознавания эмоций и генерации аватаров

2.1. Выбор датасета для задачи распознавания лицевых выражений людей

В данной работе из приведенного списка в таблице 1 был выбран наиболее современный датасет с 6 базовыми эмоциями: Real-world Affective Face Database (RAF-DB) [39] для разработки и обучения моделей и в качестве бенчмарка. Во-первых, он содержат достаточно много изображений (значительно меньше, чем AffectNet, TFD и EmotioNet, тем не менее намного больше, чем остальные) и он взят из открытых интернет источников. Во-вторых, RAF-DB полностью размечен людьми в отличие от остальных крупных наборов данных, что повышает его надежность. Он содержит цветные необработанные изображения людей в различных позах и ракурсе, с разным выражением лица, что близко к тому, как люди выглядят в реальной жизни.

Рис. 14: Примеры изображений из оригинального датасета RAF-DB [39]

2.2. Детали обучения моделей для распознавания лиц людей

2.2.1. Предобработка данных

В открытых источниках можно найти две версии датасета RAF-DB [39]: оригинальная и обработанная (aligned). В обработанной версии изображения обрезаны по лицу человека и приведены к размеру 100 на 100 пикселей. Пример оригинального и обработанного изображения можно увидеть на рисунке 15.

В данной работе использовалась обработанная версия датасета, так как она содержит меньше плохих изображений, на изображении гарантировано изображен один человек и изображения обрезаны по лицам людей, что упрощает задачу классифика-

Рис. 15: Пример обрезанного и обработанного изображения из датасета RAF-DB [39]

ции и не позволяет нейросети при обучении использовать неинформативные признаки такие как внешний фон и одежда людей.

2.2.2. EfficientNet

Для построения модели распознавания эмоций была взята современная быстрая модель EfficientNet [52], которая показывает наилучшие результаты на бенчмарке ImageNet [30]. EfficientNet версии B4 имеет почти такое же количество параметров, как ResNet50 [11] и хорошо помещается на графический процессор.

Предобученную основу(backbone) EfficientNet используем для извлечения параметров. Для распознавания эмоций была дописана ”голова” модели, которая состоит из полносвязных слоев. Более подробно архитектуру можно рассмотреть на рисунке 16.

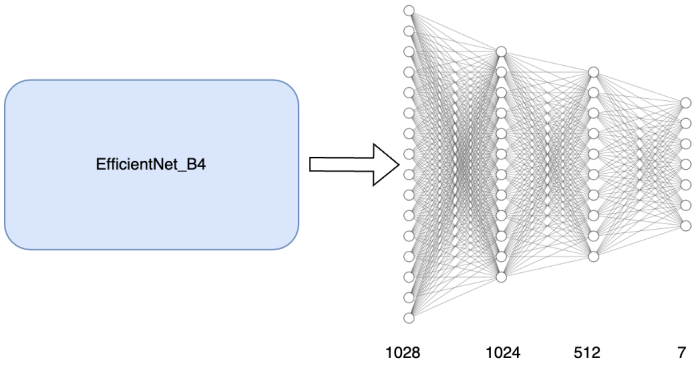


Рис. 16: Архитектура FER модели на основе EfficietNetB4 [52]

Для предобработки изображений использовались базовые аугментации:

- **SmallestMaxSize**. Изменяет размер изображения до определенного размера(388)

и сохраняет исходные пропорции

- **RandomCrop**. Обрезает изображение в случайной области по квадрату (384, 384). Изображения именно такого размера должны подаваться на вход EfficientNetB4
- **RandomBrightnessContrast**. Произвольно изменяет яркость и контрастность входного изображения
- Попиксельная нормализация изображения. Стандартная техника, которая улучшает обобщающую способность сверточных нейронных сетей

В качестве оптимизатора использовался Adam [12]. Так как мы обучаем модель путем тонкой настройки(fine-tuning), начальная скорость обучения должна выставляться маленькой - 0.00001. Для регулировки скорости обучения использовался Cosine Annealing learning rate scheduler [26]. Для рассчета функции потерь на мульти-классовой классификации использовался CrossEntropyLoss:

$$L = -\sum_{i=1}^{k} y_i log(\hat{y}_i)$$

где $k = 7$ - количество классов для нашего датасета.

Кроме того, для сравнения моделей, используя такую же архитуктуру я использовал разные основные части:

- EfficientNetV2 версии S [53]. Он является следующим поколением сети EfficientNet, легче(имеет меньшее количество обучаемых параметров), и имеет более высокий результат на бенчмарках обработки изображений.
- ResNet50 [11], который имеет такое же количество параметров, но используется в качестве основы(backbone) во многих моделях обработки изображений.

В ходе экспериментов по улучшению качества моделей, была применена новая техника аугментации изображении - TrivialAugment [48]. Это техника автоматического применения оптимальной стратегии аугментации для изображения, которая будет наиболее оптимальной для используемого датасета. Пример применения TrivialAugment на рисунке 17. Данный метод аугментации показывает хорошие результаты в задачах обработки изображений и улучшает многие современные модели.

По рисунку 18 и таблице 2 можно сделать вывод, что модели с основой от EfficientNet показывают себя лучше в данной задаче. Также, для дальнейшей работы следует использовать EfficientNetV2 в качестве предобученной модели для извлечения признаков с изображения.

Рис. 17: Обработка изображения из датасета RAF-DB [39] с помощью TrivialAugment

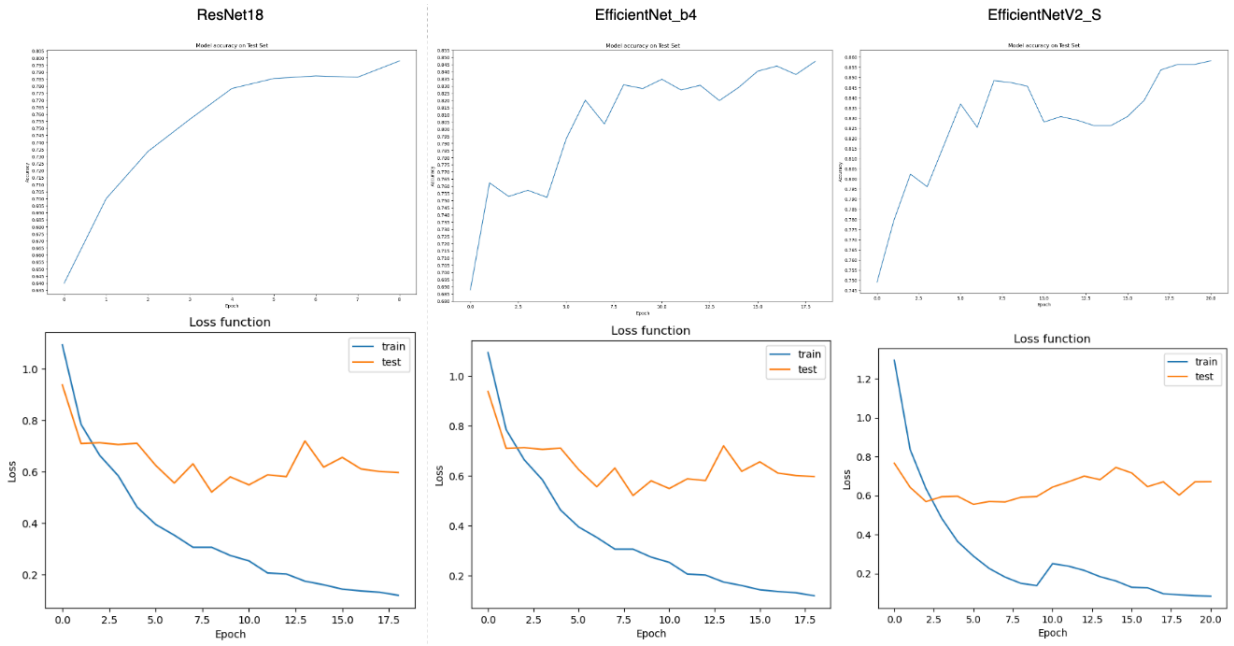


Рис. 18: Сравнение моделей FER на основе ResNet50, EfficientNet-b4, EfficientNetV2-S с применением TrivialAugment

Модель	Точность (в процентах)	Точность (в процентах) с использованием TrivialAugent
ResNet50	80.4	83.2
EfficientNet-b4	80.2	84.71
EfficientNetV2-S	83.9	85.8

Таблица 2: Сравнительная таблица результатов распознавания эмоций моделей с базовой архитектурой

2.2.3. DAN

Модель для распознавания эмоций DAN [14] показывает хорошие результаты и имеет высокие позиции в точности на бенчмарке RAF-DB [39].

Данная модель использует предобученную сверточную нейронную сеть ResNet18 [11] в качестве основы для извлечения признаков, и сеть внимания, состоящую из нескольких голов механизма внимания.

Целью эксперимента над улучшением модели было переделать архитектуру авторской модели, заменив основную часть модели ResNet18 на EfficientNetV2-S и улучшить предобработку датасета перед обучением модели.

Для замены исходного ResNet18 на EfficientNetV2-S нужно было изменить входные параметры сети механизма внимания, так как размер выходного вектора у ResNet18 равен 512, а у EfficientNetV2-S он равен 1280.

В качестве аугментаций использовались:

- Базовые авторские аугментации
- RandAugment [17]. Техника автоматического применения стратегии аугментации изображения. В настраиваемых параметрах можно указать, сколько изменений применять к изображению. Значение по умолчанию равно двум.
- TrivialAugment [48]

Для получения модели лучшего качества были проведены эксперименты с изменением гиперпараметров модели и обучения:

- Начальная скорость обучения (лучше всего работает с learning rate = 0,001)
- Размер выборки(batch) во время обучения (максимум, который помещался в память графического процессора равен 84 с использованием EfficientNetV2-S)
- Количество голов механизма внимания (варьировались от 4 до 5)
- Регулировщик скорости обучения (CosineAnnealingLR, OneCycleLR, StepLR, CosineAnnealingWarmRestarts)
- Оптимизатор (Adam [12], AdamW [27], SGD)
- Дополнительные аугментации (RandomErasion [59] - наложение темного прямоугольника случайным образом на изображение. Повышает обобщающую способность моделей и борется с переобучением)
- Замена основы модели EfficientNetV2-S на другую версию, которая имеет большее количество параметров EfficientNetV2-M. Не дало значительного улучшения качества, так как данная модель больше и во время обучения на карту графического процессора не помещались выборки с большим размером батча.

Таким образом, удалось выделить 2 самые лучшие модели, которые отображены в таблице 3 и метрики их обучения на рисунке 19.

В таблице 3 представлено 3 модели:

1. Первая модель была обучена на основе авторского кода [14] с основной предобученной моделью ResNet18 [11] с изменением размера батча при обучении, чтобы уместить на графический процессор

2. Вторая модель показала наилучший результат из всех экспериментов. В ней используется архитектура DAN с 4 головами механизма множественного внимания и предобученными EfficientNetV2-S:

   - **Аугментации:**: RandAugment [17] с 2 операциями трансформации и RandomErasing [59] для предобработки датасета
   - **Размер батча в загрузчике данных при обучении:** Максимально возможный размер 84 в графический процессор с памятью 16 ГБ
   - **Оптимизатор:** AdamW [27] с регуляризацией весов равной 0.005;
     С использованием AMSGrad [28] (метод стохастической оптимизации, направленный на устранение проблем со сходимостью оптимизаторов, которые работают по типу алгоритма Adam [12]. AMSGrad использует максмум от квадратов градиентов за прошлые эпохи, вмето экспоненциального среднего для обновления параметров оптимизатора);
     С начальной скоростью обучения 0.001
   - **Регулятор скорости обучения**: $CosineAnnealingLR(T_max = 10, eta_min = 0.00001)$ [26]
   - **Количество эпох:** Результат 88.2 процента точности был достигнут после 107 эпох обучения.

3. Третья модель построена с 5 головами механимами внимания и использует такие же характеристики, как и во второй, только для аугментаций использовался TrivialAugment [48] вместо RandAugment [17]

Метрика	DAN basic	DAN with 4 attention heads + RandAugment(2)	DAN with 5 attention heads + TrivialAugment
Точность на RAF-DB (в процентах)	87.19	88.2	87.94

Таблица 3: Сравнительная таблица результатов распознавания эмоций моделей с использованием архитектуры DAN [14]

Таким образом, использую EfficientNetV2-S [53] и архитектуру модели DAN [14] удалось получить модель распознавания эмоций с хорошей точностью.

Рис. 19: Метрики функции потерь и точности при обучении моделей на основе архитектуры DAN [14]

2.3. Генерация аватаров с помощью генеративных моделей

Для полноценной разработки и тестирования системы детектирования аномалий, необходимо получить датасет аватаров, который будет приближенно напоминать то, как будут выглядеть аватары в цифровой метавселенной. Кроме того, Необходимо сгенерировать такие изображения, которые будут качественно передавать структуру лица человека и транслировать его эмоции.

Для выполнения таких требований к сгенерированным изображениям, был выбран генератор с функцией переноса стиля на исходное изображения DualStyleGan [43].

2.3.1. Выбор целевого стиля для генерации аватаров

При выборе целевого стиля изображения в своей работе я опирался на то, чтобы сгенерированные изображения как можно более были бы похожи на реальных людей, но при этом было видно, что лицо ненастоящее и может существовать только в цифровом мире.

На рисунке 20 представлены стили, которые можно использовать с помощью DualStyleGan. Из всех представленных изображений, больше всего под поставленные требования

Рис. 20: Пример доступных стилей для переноса из модели DualStyleGan [43]

подходил анимационных стиль Pixar: сгенерированные лица людей сохраняют черты реального человека, при этом присутствует признак того, что человек из цифрового мира.

2.3.2. Технические детали генерации датасета аватаров

Для генерации изображений исползовался предобученный генератор, который заточен под стиль Pixar.

В модели есть два регулируемых параметра: Structure Weight(S) и Color Weight(C), $S, C \in [0, 1]$. Чем выше параметр S, тем глубже будет передача структуры целевого стиля на изображение, чем выше C, тем больше оно будет ближе по цветопередаче к целевому стилю.

Таким образом, можно предположить, что для генерации нашего датасета лучше будет выставить эти параметры в меньшее значение близкое к 0. Рассмотрим примеры генерации с различными значениями данных параметров на рисунке 21:

Также, есть опция выбрать энкодер для обработки изображения и перевода его в скрытое пространство: Z+ - обеспечивает более хорошую исходного изображения под целевой стиль; W+ - обеспечивает более детальную реконструкцию исходного изображения. Для нашей задачи был выбран энкодер W+.

Для того, чтобы точно оценить, какие значения параметров S и C выбрать для генерации, была проведена генерация небольшой выборки изображений (3 тысячи изображений), чтобы подать эти веса на вход обученной модели распознавания эмоций DAN 2, чтобы оценить качество передаваемых эмоций. Результат исследования

Рис. 21: Генерация изображений с различными значениями параметров Structure Weight(S) и Color Weight(C)

представлен в таблице 4.

Метрика	S=0.1,C=0.1	S=0.225,C=0.1	S=0.3,C=0.8	S=0.4,C=0.1
Точность распознавания эмоций (в процентах)	55.126	56.222	46.966	42.667

Таблица 4: Сравнительная таблица результатов распознавания эмоций модели DAN 2 на сгенерированных аватарах с разными весовыми параметрами

Таким образом, для генерации аватаров была выбрана модель DualStyleGan с энкодером W+ и параметрами Structure Weight(S) равным 0.225 и Color Weight(C) равным 0.1.

Для генерации изображений для всего датасета RAF-DB [39], который состоит из 15 тысяч изображений, потребовалось около 9 часов.

2.3.3. Примеры несоответствия эмоций исходного изображения и аватара

К сожалению, генерация такого большого датасета не оказалась полностью идеальна. На некоторых изображениях эмоции лиц людей транслировались некорректно.

На рисунке 22 изображено два примера. На первом исходном изображении изображена ярость, а на аватара этого человека транслируется удивление. На втором изображении у человека отчетлива видна эмоция грусти, а его аватар изображает счастье.

Рис. 22: Примеры неправильной трансляции эмоции

3. Система детектирования аномалий

3.1. Подготовка данных и моделей для реализации системы детектирования аномалий

Для построения системы детектирования аномалий требуются размеченные данные. Как мы заметили в прошлой главе, не все сгенерированные изображения корректно транслировали эмоции с исходного изображения.

Чтобы обучить пайплайн детекции аномалий, необходимо иметь две модели распознавания эмоций: одну для реальных лиц, а вторую для аватаров, и они обе должны иметь хорошую точность на соответствующих им изображениях.

3.1.1. Разметка датасета аватаров

Для того, чтобы дообучить хорошую FER-модель на аватарах, нужно правильно вручную разметить некоторое количество изображений из сгенерированного датасета.

В ходе ручного распознавания эмоций на аватарах, получилось собрать размечен-

ный датасет, который состоит из:

- Обучающая выборка - **958** размеченных аватаров
- Тестовая выборка - **324** размеченных аватаров

Таким образом, мы имеем датасет для дообучения модели распознавания эмоций на аватарах.

Кроме того, можем сразу сгенерировать датасет для обучения и тестирования системы детектирования аномалий.

Для этого, сравним между собой метки эмоций аватара и исходного изображения из датасета RAF-DB [39]. Из 1282 размеченных изображений, 665 аватаров корректно транслируют эмоцию из исходного изображения, остальные 617 отображают другую эмоцию. Так, получился датасет для пайплайна детекции аномалий, состоящий из 958 обучающих данных с метками:

- 0 - если аномалии нет
- 1 - если аномалия присутствует

Таким образом, мы получили размеченный датасет для обучения FER-модели и датасет для система детектирования аномалий.

3.1.2. Дообучение модели распознавания эмоций на изображениях аватаров

Рассмотрим лучшую из полученных моделей распознавания эмоций: DAN на основе EfficientNetV2-S 2.

Загрузим веса данной модели и дообучим её на датасете аватаров. Для этого, нужно предобработать изображения. Будем обрезать лица с помощью MTCNN [32]. Пример предобратки изображения на рисунке 23

Используя те же параметры, что и в исходной модели при обучении и выставив низкую скорость обучения(так как мы всего лишь дообучаем датасет на маленьком датасете), после 6 эпох удалось достигнуть **точности 78.4 процента на тестовой выборке**. Метрики обучения можно увидеть на рисунке **??**.

3.2. Разработка системы детектирования аномалий

Для построения пайплайна детектирования аномалий, я буду использовать архитектуру схожую с архитектурой Сиамской нейронной сети 1.7.1.

Рис. 23: Предобработка аватара с помощью MTCNN [32]

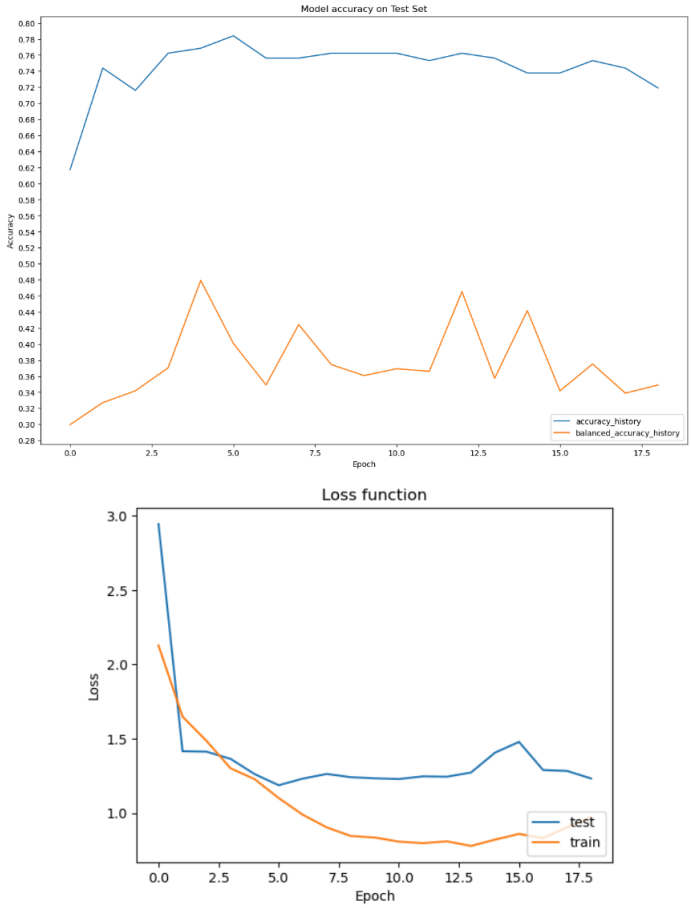


Рис. 24: Метрики обучения модели распознавания эмоций на датасете аватаров

3.2.1. Первая версия пайплайна

В данной версии будет использоваться архитектура подобно той, которая изображена на рисунке 13 и описана в главе 1.7.1.

В качестве парных моделей рассмотрим модель с точностью 88.2 процента на ре-

альных изображениях 2 и модель, дообученную на аватарах 3.1.2.

В архитектуре модели DAN [14] два последних слоя нейросети извлекают скрытые признаки из голов механизма внимания. Это два слоя:

- Линейный слой Linear, который принимает на вход вектор размером 1280 и на выходе выдает вектор размера 7 (7 базовых эмоций)
- Слой BatchNormalization

Заменим данные слои обычным Identity слоем, который просто пропускает вектор через себя, не изменяя его. Таким образом, мы получим модель, которая извлекает скрытые параметры, связанные с лицевыми выражениями, так как модели были обучены для распознавания эмоций и эти параметры можно передать в ContrastiveLoss [24].

Таким образом, обучаем пайплайн, который представлен в главе 1.7.1.

В результате обучения модели на 100 эпохах, с использованием оптимизатора Adam [12] и скорость обучения 0.0005, функция потерь уменьшится с 1115850624.0 до 1917.2569580078125, как показано на рисунке 25.

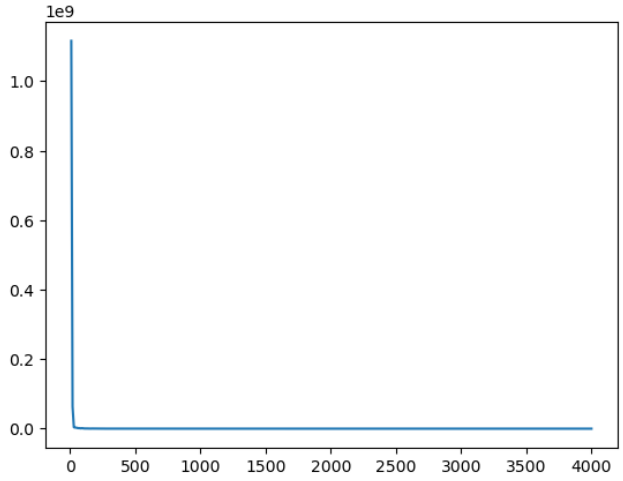


Рис. 25: Функция потери при обучении пайплайна детекции аномалий

Далее, эвристическим способом я отобрал порог по евклидову расстоянию между выходами модели, который показывает:

- Если непохожесть векторов меньше порога, то аномалии нет
- Если больше, то эмоции на изображениях отличаются

Так, при пороге 135 достигается самая высокая точность модели - 61.3 процента. На рисунке 26 можно увидеть примеры работы модели.

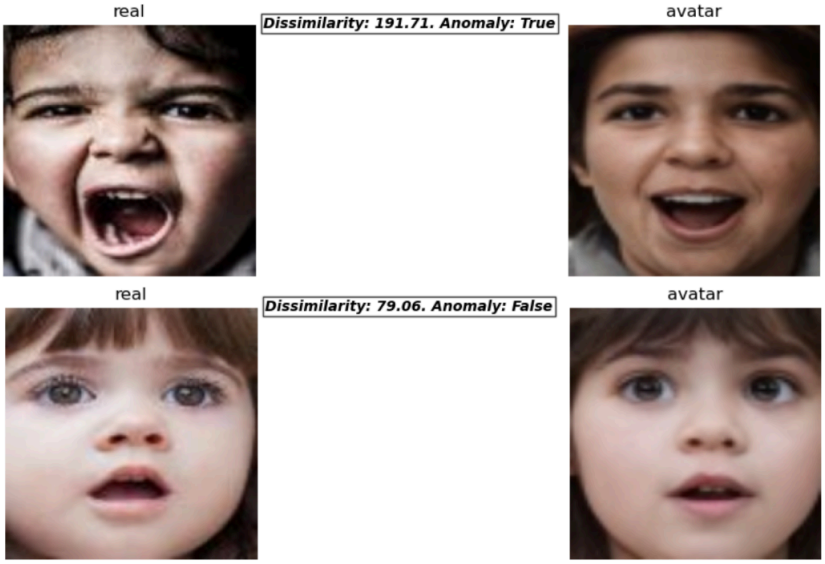


Рис. 26: Примеры выхода первой версии модели

3.2.2. Вторая версия пайплайна

В данной версии будет реализована измененная архитектура Сиамской нейронной сети.

Пусть у нас также имеются парные модели распознавания эмоций и мы удалили из них последние два слоя.

Далее, конкатенируем выходы двух моделей и напишем над ними полносвязную нейронную сеть, которая будет обрабатывать вектора и выдавать метку от 0 до 1 - есть аномалия или нет.

Технически, построим архитектуру, добавив полносвязные слои, как на рисунке 27.

В качестве функции потерь будет использоваться BinaryCrossEntropyLoss:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p(y_i)) + (1 - y_i) log(1 - p(y_i))$$

В результате обучения на 41 эпохе с использованием оптимизатора Adam [12] и скоростью обучения 0.0001 модель достигла точности в детекции 81 процент. Метрики обучения можно увидеть на рисунке 28.

Таким образом, удалось построить и обучить систему детектирования аномалий с хорошей точности нахождения несоответствий.

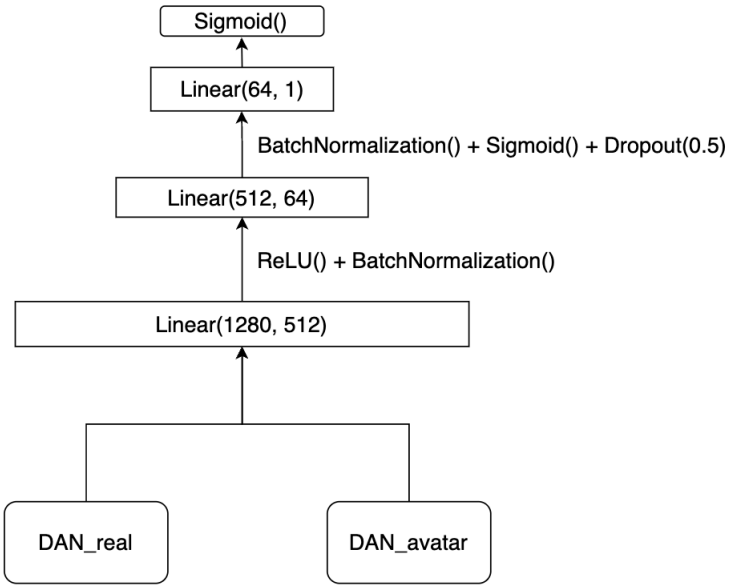


Рис. 27: Архитектура второй версии пайплайна детекции аномалий

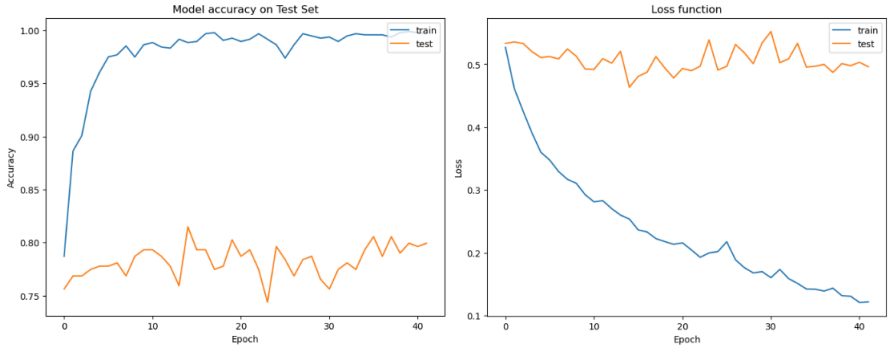


Рис. 28: Метрики обучения второй версии пайплайна детекции аномалий

Список литературы

[1] Buslaev Alexander, Iglovikov Vladimir I., Khvedchenya Eugene, Parinov Alex, Druzhinin Mikhail, and Kalinin Alexandr A. Albumentations: Fast and Flexible Image Augmentations // Information. — 2020. — feb. — Vol. 11, no. 2. — P. 125. — Access mode:

[2] Ali Mollahosseini Behzad Hasani and Mahoor Mohammad. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild // arXiv:1708.03985. — 2017.

[3] An F., Liu Z. Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM // Visual Computer. — 2020. — Vol. 36. — P. 483–498.

[4] Karras Tero, Laine Samuli, Aittala Miika, Hellsten Janne, Lehtinen Jaakko, and Aila Timo. Analyzing and Improving the Image Quality of StyleGAN. — 2020. — 1912.04958.

[5] Arjovsky Martin, Chintala Soumith, and Bottou Léon. Wasserstein GAN. — 2017. — 1701.07875.

[6] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, and Polosukhin Illia. Attention Is All You Need. — 2017. — 1706.03762.

[7] Babu D.R. , Shankar R.S., Manesh G. and Murthy K.V. Facial expression recognition using bezier curves with hausdorff distance // Proc. IEEE International Conference on IoT and Application. — 2017.

[8] Benitez-Quiroz C. Fabian, Srinivasan Ramprakash, and Martinez Aleix M. EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2016. — P. 5562–5570.

[9] Bobe A., Konyshev D. and Vorotnikov S. Emotion recognition system based on the facial motor units' analysis. — 2016. — No. 9. — P. 7.

[10] Yun Sangdoo, Han Dongyoon, Oh Seong Joon, Chun Sanghyuk, Choe Junsuk, and Yoo Youngjoon. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. — 2019. — 1905.04899.

[11] He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep Residual Learning for Image Recognition. — 2015. — 1512.03385.

[12] Diederik P. Kingma Jimmy Ba. Adam: A Method for Stochastic Optimization. — 2017. — 1412.6980.

[13] Dino H.I., Abdulrazzaq M.B. Facial expression classification based on SVM, KNN and MLP classifiers // Proc. International Conference on Advanced Science and Engineering. — 2019. — P. 70–75.

[14] Wen Zhengyao, Lin Wenzhong, Wang Tao, and Xu Ge. Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition. — 2022. — 2109.07270.

[15] Déniz O., Bueno G., Salido J. and De la Torre F. Face recognition using histograms of oriented gradients // Pattern Recognition Letters. — 2011. — Vol. 32. — P. 1598–1603.

[16] Tan Mingxing, Chen Bo, Pang Ruoming, Vasudevan Vijay, Sandler Mark, Howard Andrew, and Le Quoc V. EfficientNetV2: Smaller Models and Faster Training. — 2018. — 1807.11626.

[17] Ekin D. Cubuk Barret Zoph Jonathon Shlens Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. — 2019. — 1909.13719.

[18] Chen Junkai, Chen Zenghai, Chi Zheru, and Fu Hong. Facial Expression Recognition Based on Facial Components Detection and HOG Features. — 2014. — 07.

[19] Fan J. , Tie Y. , and L. Qi. Facial expression recognition based on multiple feature fusion in video // Proc. International Conference on Computing and Pattern Recognition. — 2018. — P. 86–92.

[20] Fu X., Fu K., Zhang Y. and Fu X. Facial expression recognition based on Curvelet transform and sparse representation // Proc. 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. — 2018. — P. 257–263.

[21] Goodfellow Ian J., Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron, and Bengio Yoshua. Generative Adversarial Networks. — 2014. — 1406.2661.

[22] Chattopadhyay Aditya, Sarkar Anirban, Howlader Prantik, and Balasubramanian Vineeth N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. — 2018. — 1710.11063.

[23] Greche L., Es-Sbai N. , Lavendelis E. Histogram of oriented gradient and multi layer feed forward neural network for facial expression identification // Proc. International Conference on Control, Automation and Diagnosis. — 2017. — P. 333–337.

[24] Hadsell Raia, Chopra Sumit, and Lecun Yann. Dimensionality Reduction by Learning an Invariant Mapping. — 2006. — 02.

[25] Huang He, Yu Philip S., and Wang Changhu. An Introduction to Image Synthesis with Generative Adversarial Nets. — 2018. — 1803.04469.

[26] Ilya Loshchilov Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. — 2016. — 1608.03983.

[27] Ilya Loshchilov Frank Hutter. Decoupled Weight Decay Regularization. — 2019. — 1711.05101.

[28] Ilya Loshchilov Frank Hutter. On the Convergence of Adam and Beyond. — 2019. — 1904.09237.

[29] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. — 2021. — 2010.11929.

[30] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, Berg Alexander C., and Fei-Fei Li. ImageNet Large Scale Visual Recognition Challenge. — 2015. — 1409.0575.

[31] Iqtait M., Mohamad F.S. and Mamat M. Feature extraction for face recognition via active shape model (ASM) and active appearance model (AAM) // IOP Conference Series: Materials Science and Engineering. — 2018. — Vol. 332.

[32] Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, and Qiao Yu. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks // IEEE Signal Processing Letters. — 2016. — oct. — Vol. 23, no. 10. — P. 1499–1503. — Access mode:

[33] Jumani S.Z., Ali F., Guriro S., Kandhro I.A., Khan A., Zaidi A. Facial expression recognition with histogram of oriented gradients using CNN // Indian Journal of Science and Technology. — 2019. — P. 1–8.

[34] Karras Tero, Laine Samuli, and Aila Timo. A Style-Based Generator Architecture for Generative Adversarial Networks. — 2019. — 1812.04948.

[35] Zhang Yuhang, Wang Chengrui, Ling Xu, and Deng Weihong. Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition. — 2022. — 2207.10299.

[36] Li S. , Gong D. , and Y. Yuan. Face recognition using Weber local descriptors // Neurocomputting. — 2013. — Vol. 122. — P. 272–283.

[37] Li Jia. Facial Expression Recognition using Vanilla ViT backbones with MAE Pretraining. — 2023. — 2207.11081.

[38] Li Shan and Deng Weihong. Deep Facial Expression Recognition: A Survey // arXiv. — 2018.

[39] Li Shan and Deng Weihong. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition // IEEE Transactions on Image Processing. — 2019. — Vol. 28, no. 1. — P. 356–370.

[40] Guo Yandong, Zhang Lei, Hu Yuxiao, He Xiaodong, and Gao Jianfeng. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. — 2016. — 1607.08221.

[41] Mirza Mehdi and Osindero Simon. Conditional Generative Adversarial Nets. — 2014. — 1411.1784.

[42] Mustafa A., Oulefki A., Bengherabi M., Boutellaa E. and Algaet M. Towards nonuniform illumination face enhancement via adaptive contrast stretching // Multimedia Tools and Applications. — 2017. — Vol. 76. — P. 21961–21999.

[43] Yang Shuai, Jiang Liming, Liu Ziwei, and Loy Chen Change. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. — 2022. — 2203.13248.

[44] Psaroudakis Andreas and Kollias Dimitrios. MixAugment Mixup: Augmentation Methods for Facial Expression Recognition. — 2022. — 2205.04442.

[45] Xu Jing, Pan Yu, Pan Xinglin, Hoi Steven, Yi Zhang, and Xu Zenglin. RegNet: Self-Regulated Network for Image Classification. — 2021. — 2101.00590.

[46] Ryumina E. and Karpov A. Analytical review of methods for emotion recognition by human face expressions // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. — 2020. — Vol. 20. — P. 163–176.

[47] Liu Wei, Anguelov Dragomir, Erhan Dumitru, Szegedy Christian, Reed Scott, Fu Cheng-Yang, and Berg Alexander C. SSD: Single Shot MultiBox Detector // Computer Vision – ECCV 2016. — Springer International Publishing, 2016. — P. 21–37. — Access mode:

[48] Samuel G. Müller Frank Hutter. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation. — 2021. — 2103.10158.

[49] Shan C., Gong S., and P.W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study // Image and Vision Computing. — 2009. — Vol. 27. — P. 803–816.

[50] Liu Ze, Lin Yutong, Cao Yue, Hu Han, Wei Yixuan, Zhang Zheng, Lin Stephen, and Guo Baining. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. — 2021. — 2103.14030.

[51] Talegaonkar I., Joshi K., Valunj S., Kohok R., Kulkarni A. Real time facial expression recognition using deep learning // Proc. of International Conference on Communication and Information Processing. — 2019.

[52] Tan Mingxing and Le Quoc V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. — 2020. — 1905.11946.

[53] Tan Mingxing and V. Quoc. EfficientNetV2: Smaller Models and Faster Training. — 2021. — 2104.00298.

[54] Tripathi A. and Pandey S. Efficient facial expression recognition system based on geometric features using neural network // Lecture notes in networks and systems. — 2018. — P. 181–190.

[55] Varma S., Shinde M. , Chavan S.S.,. Analysis of PCA and LDA features for facial expression recognition using SVM and HMM classifiers // Techno-Societal Proc. 2nd International Conference on Advanced Technologies for Social Applications. — 2020. — Vol. 1. — P. 109–119.

[56] Viola P. and Jones M.J. Robust real-time face detection // International Journal of Computer Vision. — 2004. — Vol. 57. — P. 137–154.

[57] Wang X. and Chen L. Contrast enhancement using feature-preserving bi-histogram equalization // Signal Image and Video Processing. — 2018. — Vol. 12. — P. 685–692.

[58] Zhao J., Mao X., Zhang J. Learning deep facial expression features from image and optical flow sequences using 3D CNN // Visual Computer. — 2018. — Vol. 34. — P. 1461–1475.

[59] Zhun Zhong† Liang Zheng Guoliang Kang Shaozi Li Yi Yang. Random Erasing Data Augmentation. — 2017. — 1708.04896.

Приложение