Рабчевский, А. Н. Обзор методов и систем генерации синтетических обучающих данных / А.Н. Рабчевский // Прикладная математика и вопросы управления. – 2023. – № 4. – С. 6–45. DOI 10.15593/2499-9873/2023.4.01

Библиографическое описание согласно ГОСТ Р 7.0.100–2018
Рабчевский, А. Н. Обзор методов и систем генерации синтетических обучающих данных / А.Н. Рабчевский. – Текст : непосредственный. – DOI 10.15593/2499-9873/2023.4.04 // Прикладная математика и вопросы управления / Applied Mathematics and Control Sciences. – 2023. – № 4. – С. 6–45.

пермский политех

ПРИКЛАДНАЯ МАТЕМАТИКА И ВОПРОСЫ УПРАВЛЕНИЯ
№ 4, 2023
https://ered.pstu.ru/index.php/amcs

Научный обзор
DOI: 10.15593/2499-9873/2023.4.01
УДК 004.855.5

Обзор методов и систем генерации синтетических обучающих данных

А.Н. Рабчевский

Пермский государственный национальный исследовательский университет, Пермь, Российская Федерация
ООО «СЕУСЛАБ», Пермь, Российская Федерация

О СТАТЬЕ

Получена: 07 февраля 2023
Одобрена: 28 сентября 2023
Принята к публикации: 14 декабря 2023

Финансирование
Исследование не имело спонсорской поддержки.
Конфликт интересов
Автор заявляет об отсутствии конфликта интересов.
Вклад автора
100 %.

Ключевые слова:
искусственный интеллект, нейросетевые алгоритмы, синтетические данные, генерация, сверточные нейронные сети, CNN, глубокие нейронные сети, DNN, генеративные состязательные сети, GAN, адаптация домена, рандомизация, 3D-модели, симуляция, виртуальная реальность.

АННОТАЦИЯ

Развитие современных систем искусственного интеллекта невозможно себе представить без нейросетевых технологий. В процессе проектирования исследователи часто сталкиваются с тем, что данных для обучения современных нейросетевых моделей недостаточно, эти данные могут быть несбалансированны или сильно разрежены. Нередко случается, что реальных данных просто не существует, так как область исследований еще только формируется. Актуальной является проблема обеспечения конфиденциальности реальных персональных данных или медицинских данных пациентов, которые используются в процессе обмена между исследователями или в процессе тестирования различных нейросетевых систем. Во многих предметных областях стоимость сбора и разметки реальных данных может быть чрезмерно высокой. Для решения этих проблем все чаще используются синтетические данные.

Цель данной публикации состоит в том, чтобы познакомить читателей с достижениями в области генерации и использования обучающих синтетических данных. В работе представлено описание различных методов, систем и программных средств, используемых для генерации синтетических данных, которые могут помочь в улучшении нейросетевых моделей. Поскольку в настоящее время уже сформировалась целая индустрия по производству синтетических данных, представлены ведущие технологические платформы синтеза данных. Работа носит обзорный характер, поэтому содержит обширную библиографию. Ценность статьи заключается в том, что приведенный обзор поможет читателям расширить представления об использовании синтетических данных в решении широкого спектра нейросетевых задач, а также глубже познакомится с методами и инструментами для их генерации.

© **Рабчевский Андрей Николаевич** – кандидат технических наук, старший преподаватель, кафедра информационной безопасности и систем связи; заместитель директора по науке, e-mail: ran@psu.ru, ORCID 0000-0002-4096-9145.

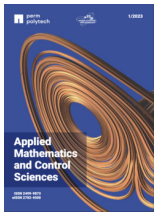
Perm Polytech Style: Rabchevsky A.N. Review of methods and systems for generation of synthetic training data. *Applied Mathematics and Control Sciences*, 2023, no. 4, pp. 6–45. DOI: 10.15593/2499-9873/2023.4.01

MDPI and ACS Style: Rabchevsky, A.N. Review of methods and systems for generation of synthetic training data.. *Appl. Math. Control Sci.* **2023**, 4, 6–45. https://doi.org/10.15593/2499-9873/2023.4.01

Chicago/Turabian Style: Rabchevsky, Andrey N. 2023. Review of methods and systems for generation of synthetic training data". *Appl. Math. Control Sci.* no. 4: 6–45. https://doi.org/10.15593/2499-9873/2023.4.01

perm polytech

APPLIED MATHEMATICS AND CONTROL SCIENCES
№ 4, 2023
https://ered.pstu.ru/index.php/amcs



Review
DOI: 10.15593/2499-9873/2023.4.01
UDK 004.855.5

Review of methods and systems for generation of synthetic training data

A.N. Rabchevsky

Perm State University, Perm, Russian Federation
LLC "SEUSLAB", Perm, Russian Federation

ARTICLE INFO

Received: 07 February 2023
Approved: 28 September 2023
Accepted for publication:
14 December 2023

Funding
This research received no external funding.
Conflicts of Interest
The author declare no conflict of interest.
Author Contributions
100 %.

Keywords:
artificial intelligence, neural network algorithms, synthetic data, generation, convolutional neural networks, CNN, deep neural networks, DNN, generative adversarial networks, GAN, domain adaptation, randomization, 3D models, simulation, virtual reality.

ABSTRACT

It is impossible to imagine the advancement of modern artificial intelligence systems without neural network technologies. During the design process researchers are often faced with the fact that there is not enough data to train modern neural network models, these data may be unbalanced or highly sparse. Often it happens that real data simply does not exist, as the research field is still emerging. A relevant problem is ensuring the confidentiality of real personal or patient medical data, which is used in the exchange between researchers or in the testing of various neural network systems. In many subject areas, the cost of collecting and marking up real data can be very high. Synthetic data is increasingly being used to solve these problems.

The purpose of this publication is to introduce readers to advances in the generation and use of synthetic data. The paper presents a description of various methods, systems and software tools used to generate synthetic data, which can help to improve neural network models. Since an entire industry for synthetic data production has already formed, the leading data synthesis technology platforms are presented. The paper is of an overview nature, so it contains an extensive bibliography. The value of the article lies in the fact that this review will help readers broaden their understanding of the use of synthetic data in solving a wide range of neural network problems, as well as to become more familiar with the methods and tools for their generation.

© **Andrey N. Rabchevsky** – CSc of Technical Sciences, Senior Lecturer, Department of Information Security and Communication Systems; Deputy Director for Science, e-mail: ran@psu.ru, ORCID: 0000-0002-4096-9145.

Введение

В процессе проектирования современных нейросетевых моделей исследователи часто сталкиваются с проблемой недоступности достаточного количества данных для их обучения, а также с неравномерностью или разреженностью этих данных. Нередко случается, что реальных данных просто не существует, так как область исследований еще только формируется. Также существует проблема конфиденциальности реальных персональных данных или медицинских данных пациентов, которые используются в процессе обмена между исследователями или в процессе тестирования различных нейросетевых систем. Во всех этих случаях на помощь приходят синтетические данные.

Как следует из названия, синтетические данные, это данные, которые созданы искусственно, а не в результате реальных событий. Они часто создаются с помощью алгоритмов и используются для широкого спектра действий.

Одно из первых упоминаний о применении синтетических данных встречается в связи с разработкой и тестированием системы обнаружения вторжений агентством DARPA[1] в 1998 и 1999 гг. [1]. Тестовые данные содержали сетевой трафик и файлы журнала системных вызовов из смоделированной большой компьютерной сети. Атакующие данные были сгенерированы синтетически на основе сценариев возможных атак, а фоновые данные – с помощью программных автоматов, имитирующих использование различных услуг. Использование синтетических данных позволило разработчикам смоделировать и протестировать различные сценарии вторжения, которые ранее еще не встречались.

В тех случаях, когда новые службы еще только тестируются перед вводом в эксплуатацию, данные для обучения нейросетей могут просто отсутствовать. В этом случае для тестирования нужны синтезированные данные. Например, авторы статьи [2] уже в 2002 г. применяли синтетические данные при создании системы детектирования мошенничества.

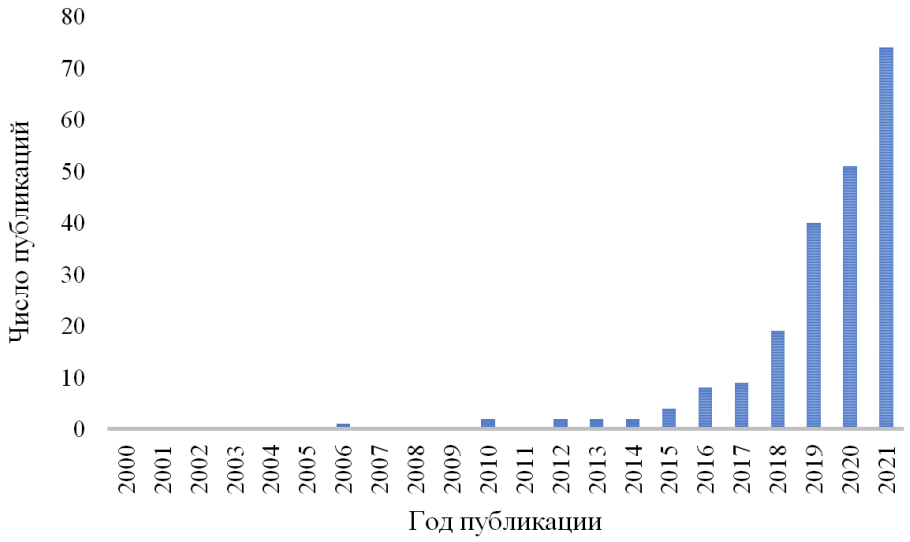


Рис. 1. Ежегодное количество публикаций, содержащих в названии термин «синтетические данные»

В последние годы использование синтетических данных стало применяться все чаще. На рис. 1 показан график роста ежегодного количества публикаций, содержащих в назва-

[1] Управление Министерства обороны США, отвечающее за разработку новых технологий для использования в интересах вооруженных сил (Defense Advanced Research Projects Agency).

нии термин «синтетические данные», опубликованных в электронной библиотеке arxiv.org[2] за период 2000–2021 гг. За первую половину 2022 г. в этой библиотеке уже опубликовано столько же статей на эту тему, сколько за весь 2021 г., и количество подобных публикаций продолжает расти.

Поиск по всем возможным источникам, вероятно, может дать еще более впечатляющую картину. На рис. 2 показан график роста количества публикаций в разрезе различных отраслей науки, также полученный на основе данных библиотеки Корнеллского университета.

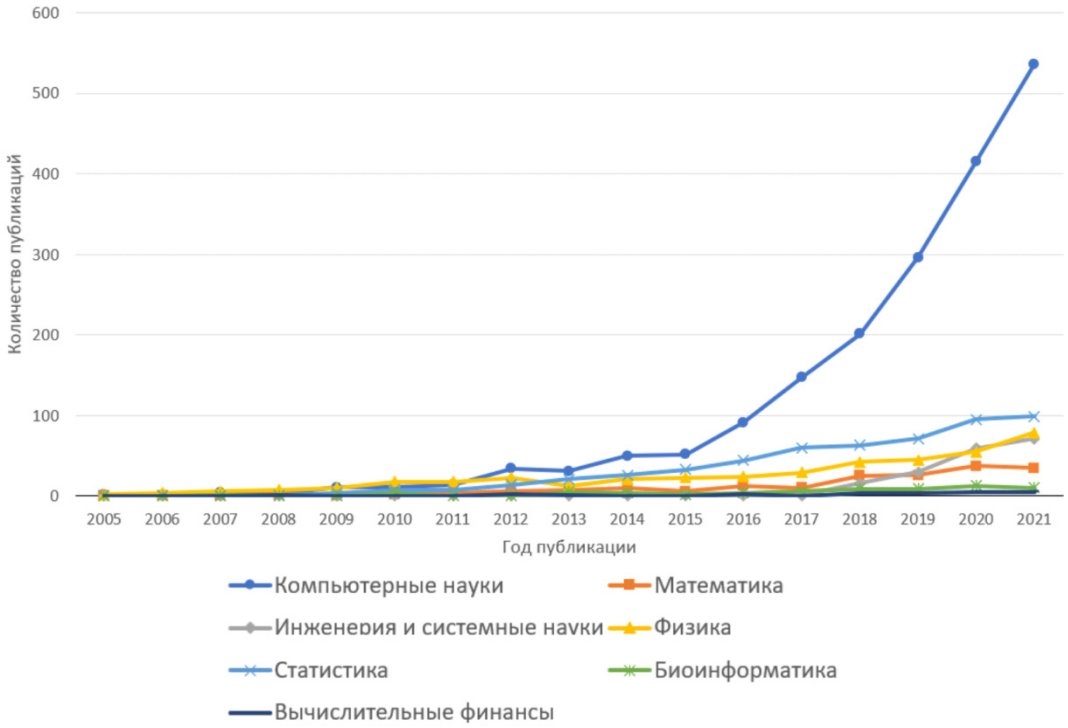


Рис. 2. Ежегодное количество публикаций с упоминанием термина «синтетические данные» по отраслям науки

Как видно из графика, основной рост публикаций наблюдается в компьютерных науках, основной вклад в которые вносят области знаний, связанные с распознаванием образов и компьютерным зрением. Наряду с бурным ростом количества публикаций наблюдается также и расширение областей применения синтетических данных. Так, по оценкам Gartner[3] к 2030 г. количество синтетических данных для обучения нейросетевых моделей превысит количество реальных данных.

Таким образом, одной из важнейших задач развития нейросетевых алгоритмов становится создание и совершенствование алгоритмов и методов генерации синтетических данных. В настоящей работе была предпринята попытка выполнить обзор существующих методов и систем генерации синтетических данных на основе анализа открытых публикаций за период с 2000 г. по первую половину 2022 г. Для генерации синтетических данных используются различные методы, которые условно можно разделить по основным принципам, используемым

[2] Электронная библиотека университета Cornell University [Электронный ресурс]. – URL: https://arxiv.org (дата обращения: 26.03.2023)

[3] Linden A. Is Synthetic Data the Future of AI? [Электронный ресурс] // Gartner. – URL: https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai (дата обращения: 26.03.2023).

в механизме генерации. Именно в таком ключе представлены методы генерации обучающих синтетических данных в начале обзора. Далее предложено описание наиболее известных систем и коммерческих платформ, используемых для генерации синтетических данных.

Общие подходы к созданию синтетических данных

В данном разделе представлены методы, которые в своей основе имеют некие общие или комплексные подходы к механизму генерации, поэтому их трудно отнести к той или иной группе. Например, в процессе создания системы детектирования мошенничества был разработан метод генерации синтетических данных, полученных на основе достоверных данных. В статье [2] авторы определяли важные характеристики достоверных данных и данных мошенничества, которые хотели обнаружить, и генерировали синтетические данные с этими свойствами. В приложении к статье размещен текст докторской диссертации[4] одного из соавторов, в главах 4.6 – 4.7 которой представлено детальное описание системы генерации синтетических данных.

В работе [3] авторами разработан комплексный метод синтеза обучающих данных, которые использовались для выявления свистов зубатых китов Odontoceti из записей гидрофонов. Для генерации данных образцы аннотированных человеком-аналитиком образцов свиста (DCLDE 2011) вводились в спектрограммы данных о звуках океана. Таким образом были созданы бинарные маски меток для каждого синтетического примера («1» указывает на наличие свиста), а также выполнена свертка масок со случайным гауссовым фильтром, чтобы размыть каждую бинарную маску. Далее размытые бинарные контурные маски совмещались с участком спектрограммы, на котором свист отсутствовал, и получались обучающие образцы со свистом.

Электронные медицинские карты (ЭМК) являются основной составляющей медицинской документации, однако обмен ЭМК или использование их для тестирования нового программного обеспечения связан с риском утечки конфиденциальных медицинских данных пациентов. В работах [4–6] описываются алгоритмы и методы, с помощью которых генерировались синтетические данные для создания синтетических ЭМК. Так, в работе [4] синтетические ЭМК создавались путем добавления модели оказания медицинских услуг в ответ на основные жалобы пациентов. Авторы статьи [5] использовали реальные электронные медицинские карты в качестве модели для создания синтетических электронных карт. В работе [6] метод создания синтетических ЭМК состоит из трех основных этапов: 1) идентификация синтетического пациента и генерация основной информации; 2) определение моделей ухода, которые получат синтетические пациенты, на основе информации, присутствующей в реальных электронных медицинских картах для аналогичных проблем со здоровьем; 3) адаптация этих моделей ухода к популяции синтетических пациентов.

В статье [7] представлен интересный подход к синтезу обучающих данных, содержащих только один пример каждого класса. Авторы предлагают так называемый жадный алгоритм коррекции смещения. Вместо того, чтобы учиться на большом, но фиксированном наборе примеров, весь обучающий набор генерируется с использованием только синтетических примеров. Цель авторов состоит в том, чтобы обучить классификатор, который обобщается на несинтетическую область без предварительного обучения или тонкой настройки на реальных данных.

[4] Kvarnström H. On the implementation and protection of fraud detection systems: Ph. D. thesis. – Göteborg, Sweden [CHALMERS UNIVERSITY OF TECHNOLOGY], 2004. – 210 p.

Дифференциальная приватность или конфиденциальность[5] (DP) – это набор систем и рекомендаций, которые помогают обеспечить безопасность и конфиденциальность данных частных лиц. Авторы статьи [8] предлагают общий подход к созданию дифференциально частных синтетических данных, состоящий из трех этапов: 1) выбор коллекции низкоразмерных краевых показателей; 2) измерение этих краевых значений с помощью механизма добавления шума; 3) генерирование синтетических данных, которые хорошо сохраняют измеренные краевые значения. Центральное место в этом подходе занимает Private-PGM, метод постобработки, который используется для оценки высокоразмерного распределения данных по зашумленным измерениям его краевых значений. Они представляют два механизма, NIST-MST и MST, которые являются примерами этого общего подхода. NIST-MST стал победителем в конкурсе синтетических данных NIST 2018 г. по дифференциальной конфиденциальности, а MST – это новый механизм, который может работать в более общих условиях и при этом демонстрирует сравнимые с NIST-MST результаты. Авторы считают, что представленный общий подход должен представлять широкий интерес и может быть использован в будущих механизмах для генерации синтетических данных.

Проблема повышения безопасности синтетических данных в условиях дифференциальной приватности решается в работе [9] путем объединения в конвейер NA+MI методов анализа синтетических данных из области множественной интерполяции и генерации синтетических данных с помощью байесовского моделирования с учетом шумов, который позволяет вычислять точные оценки неопределенности для величин популяционного уровня на основе синтетических данных DP. Чтобы реализовать NA+MI для генерации дискретных данных из граничных запросов, авторы разработали новый алгоритм генерации синтетических данных NAPSU-MQ с учетом шума, используя принцип максимальной энтропии. Дальнейшие эксперименты показали, что этот конвейер способен создавать точные доверительные интервалы из синтетических данных DP. Интервалы становятся шире при более строгой конфиденциальности, чтобы точно отразить дополнительную неопределенность, возникающую из-за шума DP.

Распространенным подходом к синтетическим данным является выборка из подогнанной модели. В работе [10] показано, что при общих предположениях такой подход приводит к выборке с неэффективными оценками, а совместное распределение выборки не согласуется с истинным распределением. Мотивируя это, авторы предлагают общий метод получения синтетических данных, который широко применим для параметрических моделей, а также имеет асимптотически эффективную сводную статистику, легко реализуем и эффективен с вычислительной точки зрения. Такой подход позволяет строить как частично синтетические наборы данных, которые сохраняют определенные итоговые статистики, так и полностью синтетические данные, которые удовлетворяют дифференциальной конфиденциальности. Авторы доказывают, что в случае непрерывных случайных величин их метод сохраняет эффективную оценку с асимптотически пренебрежимо малой ошибкой, и показывают с помощью моделирования, что это свойство выполняется и для дискретных распределений.

Федеративное обучение позволяет обучать модель, используя данные на нескольких клиентах без передачи исходных данных. Однако стандартный метод заключается в передаче параметров модели (или обновлений), которые для современных нейронных сетей могут составлять миллионы элементов, что влечет за собой значительные вычислительные затраты для клиентов. Авторы работы [11] предлагают метод федеративного обучения, при котором вместо того, чтобы передавать на сервер обновление градиента, они передают небольшое количе-

[5] Дифференциальная конфиденциальность – это набор систем и рекомендаций, которые помогают обеспечить безопасность и конфиденциальность данных частных лиц.

ство синтетических «данных». Описана процедура и показаны некоторые экспериментальные результаты, свидетельствующие, что эта процедура имеет потенциал, обеспечивая более чем на порядок снижение коммуникационных затрат при минимальной деградации модели.

Для снижения затрат на связь в федеративном обучении с большими моделями весьма актуальной задачей является сжатие модели. В работе [12] предложена схема, в которой вместо передачи обновления модели каждый клиент изучает и передает легкий синтетический набор данных, чтобы, используя его в качестве обучающих данных, модель работала одинаково хорошо на реальных обучающих данных. Сервер восстанавливает локальное обновление модели через синтетические данные и применяет стандартное агрегирование. Авторы предложили новый алгоритм FedSynth для локального обучения синтетических данных.

Обучающие множества, основанные на реальных данных, часто бывают сильно перекошены, когда одни классы значительно превосходят другие, поэтому при прогнозировании этих недостающих экземпляров алгоритмы машинного обучения не могут достичь значительной эффективности. Для решения проблемы перекоса данных было предложено множество вариантов методов синтетической избыточной выборки меньшинств (SMOTE) для балансировки наборов данных с непрерывными признаками. Однако для смешанных наборов данных с номинальными и непрерывными признаками единственным методом перевыборки на основе SMOTE, позволяющим сбалансировать данные, является метод SMOTE-NC [13]. В статье [14] авторы представляют новый метод перевыборки меньшинств, SMOTE-ENC (SMOTE – Encoded Nominal and Continuous), в котором номинальные признаки кодируются как числовые значения, а разница между двумя такими числовыми значениями отражает степень изменения ассоциации с классом меньшинств. Эксперименты показали, что модель классификации, использующая метод SMOTE-ENC, предсказывает лучше, чем модель, использующая SMOTE-NC, когда набор данных содержит значительное количество номинальных признаков, а также когда существует некоторая связь между категориальными признаками и целевым классом. Кроме того, предложенный метод устранил одно из основных ограничений алгоритма SMOTE-NC, а именно применимость только к смешанным наборам данных, которые имеют признаки, состоящие как из непрерывных, так и номинальных признаков, и невозможность работать, если все признаки набора данных являются номинальными. Метод SMOTE-ENC пригоден для применения как на смешанных наборах данных, так и на наборах, состоящих только из номинальных признаков.

В статье [15] также рассматривается проблема обработки несбалансированных данных в задаче классификации с несколькими признаками. Проблема решается с помощью двух новых методов, которые, в первую очередь, используют геометрические отношения между векторами признаков. Первый из них – алгоритм недостаточной выборки, использующий угол между векторами признаков для отбора более информативных образцов, отбрасывая при этом менее информативные. Предлагается подходящий критерий для определения информативности данной выборки. Второй – алгоритм чрезмерной выборки, который использует генеративный алгоритм для создания новых синтетических данных с соблюдением всех границ классов. Это достигается путем нахождения «ничейной земли» на основе евклидова расстояния между векторами признаков. Эффективность предложенных методов анализируется на примере решения общей задачи многоклассового классификатора на основе смеси гауссиан. Превосходство предложенных алгоритмов установлено путем сравнения с другими современными методами, включая SMOTE и ADASYN[6] (репозито-

[6] ADASYN [Электронный ресурс]. – URL: https://github.com/stavskal/ADASYN (дата обращения: 29.03.2023).

рий, где хранится программный код с модулем Python, который реализует технику адаптивной передискретизации для перекошенных наборов данных), на десяти различных общедоступных наборах данных, демонстрирующих высокий и экстремальный дисбаланс данных. Эти два метода объединены в единую систему обработки данных и названы GICaPS, чтобы подчеркнуть роль выборки информации на основе геометрии (GI) и синтеза с приоритетами классов (CaPS) в решении проблемы дисбаланса данных в нескольких классах, тем самым внося новый вклад в эту область.

Одной из самых больших проблем, осложняющих прикладное контролируемое машинное обучение, является необходимость в огромных объемах маркированных данных. Активное обучение (AL) – хорошо известный стандартный метод эффективного получения маркированных данных, когда сначала маркируются образцы, содержащие наибольшее количество информации, на основе стратегии запроса. Несмотря на то, что в прошлом было предложено множество методов для стратегий запросов, до сих пор не найдено однозначного превосходного метода, который хорошо работает в целом для всех областей. Кроме того, многие стратегии требуют больших вычислительных затрат, что препятствует широкому использованию AL для крупномасштабных проектов аннотирования. Поэтому в работе [16] предлагается IMITAL, новая стратегия запросов, которая рассматривает AL как задачу обучения ранжированию. Для обучения базовой нейронной сети авторы выбрали метод имитационного обучения. Необходимый для обучения наглядный экспертный опыт генерируется на основе чисто синтетических данных. Чтобы показать общую и превосходную применимость IMITAL, они проводят обширную оценку, сравнивая эту стратегию на 15 различных наборах данных из широкого спектра областей с 10 различными современными стратегиями запросов. Они также показывают, что их подход более эффективен во время выполнения, чем большинство других стратегий, особенно на очень больших наборах данных.

Большой успех машинного обучения на огромных массивах данных достигается ценой огромных вычислительных затрат и хранения данных для обучения и настройки. Однако существующие подходы имеют фундаментальные ограничения в оптимизации из-за ограниченной репрезентативности синтетических наборов данных без учета каких-либо характеристик регулярности данных. В связи с этим в работе [17] предлагается новая система сжатия, которая генерирует множество синтетических данных с ограниченным бюджетом на хранение с помощью эффективной параметризации с учетом регулярности данных. Далее авторы предлагают унифицированный алгоритм, который значительно улучшает качество сжатых данных по сравнению с текущим современным уровнем на CIFAR-100[7], ImageNet[8] и Speech Commands[9].

[7] Целью данного проекта является классификация изображений, маркированных по 100 категориям. Набор данных CIFAR 100M используется для обучения сверточных нейронных сетей [Электронный ресурс]. – URL: https://github.com/Sripriya07/CIFAR-100 (дата обращения: 28.03.2023).

[8] ImageNet – это база данных изображений, организованная в соответствии с иерархией WordNet (в настоящее время только существительные), в которой каждый узел иерархии представлен сотнями и тысячами изображений. Проект сыграл важную роль в развитии исследований в области компьютерного зрения и глубокого обучения [Электронный ресурс]. – URL: https://image-net.org/download.php (дата обращения: 05.07.2023).

[9] Speech Commands – набор односекундных аудиофайлов в формате .wav, каждый из которых содержит одно произнесенное английское слово. Эти слова взяты из небольшого набора команд и произнесены разными дикторами. Аудиофайлы организованы в папки по содержащимся в них словам, и этот набор данных предназначен для обучения простых моделей машинного обучения [Электронный ресурс]. – URL: https://github.com/sankalp2610/Speech_Command_Recognition (дата обращения: 28.03.2023).

Утверждая, что сеть, обученная на синтетических данных, относительно плохо работает на реальных изображениях, в работе [18] авторы предлагают принципиально иной способ работы с синтетическими изображениями, который не требует наличия реальных изображений во время обучения. В основе лежит наблюдение, что классы переднего и заднего плана не одинаково подвержены влиянию изменения домена и поэтому должны обрабатываться по-разному. То есть классы переднего плана должны обрабатываться на основе обнаружения, чтобы лучше учесть тот факт, что их форма выглядит естественно, в то время как на синтетических изображениях их текстура не является фотореалистичной. Проведенные эксперименты подтверждают эффективность такого подхода на городских пейзажах и CamVid с моделями, обученными только на синтетических данных.

Обработка, передача и хранение данных – ключевая проблема, возникающая при работе с большими массивами данных. В работе [19] представлена схема для генерации больших синтетических наборов данных «на лету», подходящих для этих методов анализа данных, которая является одновременно вычислительно эффективной и применимой к различным проблемам. Представлен пример применения предложенной схемы, а также математический анализ ее вычислительной эффективности, демонстрирующий ее эффективность.

Автоматическое определение жизненно важных признаков на видео, таких как оценка частоты сердечных сокращений и дыхания, является сложной исследовательской задачей в области компьютерного зрения, имеющей важное применение в медицине. Одной из основных трудностей при решении этой задачи является отсутствие достаточного количества контролируемых обучающих данных, что сильно ограничивает использование мощных глубоких нейронных сетей. В работе [20] эта проблема решается с помощью нового подхода глубокого обучения, в котором рекуррентная глубокая нейронная сеть обучается определять жизненно важные признаки в инфракрасной тепловой области на основе чисто синтетических данных. Самое удивительное, что данный метод создания синтетических обучающих данных является общим, относительно простым и практически не требует предварительных знаний в области медицины. Более того, система, которая обучается чисто автоматически и не нуждается в аннотации человека, также учится предсказывать дыхание или интенсивность сердцебиения для каждого момента времени и определять область интереса, которая наиболее релевантна для данной задачи, например, область носа в случае дыхания.

Для решения проблем, связанных с ограниченностью данных и нехваткой меток в имеющихся данных, авторы статьи [21] предлагают генерировать условные синтетические данные, которые будут использоваться наряду с реальными данными для разработки надежных ML-моделей. В работе была представлена гибридная модель, состоящая из условного генеративного потока и классификатора для условной генерации синтетических данных. Классификатор определяет представление признаков для условия, которое подается в поток для извлечения локального шума. Авторы генерируют синтетические данные, манипулируя локальным шумом с фиксированным условным представлением признаков. Они также предложили полуконтролируемый подход для создания синтетических образцов при отсутствии меток для большинства доступных данных. В результате была выполнена условная генерация синтетических образцов для компьютерной томографии (КТ) грудной клетки, соответствующих нормальным, с заболеванием COVID-19 и пораженным пневмонией пациентам. Показано, что предложенный метод значительно превосходит существующие модели как по качественным, так и по количественным показателям, а полуконтролируемый подход может эффективно синтезировать условные образцы в условиях нехватки меток. В качестве примера последующего использования синтетических данных

показано улучшение обнаружения COVID-19 на компьютерных томограммах с помощью условного дополнения синтетических данных.

Как видим, в данном разделе представлены общие методы, основанные на различных подходах к использованию реальных данных и смешиванию их с синтетическими для получения наилучшего результата для решения различных нейросетевых задач. В связи с таким многообразием методов невозможно точно сформулировать, какой класс методов применяется для тех или иных задач. В то же время очевидно, что каждая задача решалась уникальным методом, но все эти методы объединяет то, что направлены они на генерацию синтетических данных.

Методы генерации синтетических данных на основе рандомизации

Рандомизация, или использование генераторов случайных чисел часто применяется для создания синтетических данных. Самый простой пример использования генератора случайных чисел представлен в работах [22–24], где данные для обучения классификатора ролей пользователей социальных сетей были сгенерированы с использованием генератора случайных чисел в пределах диапазонов, заданных экспертом в данной предметной области.

Более сложный вариант использования рандомизации был применен при решении проблемы извлечения манипулятором робота объекта из кучи мусора. Проблема связана с точным определением границ объекта и той стороны, за которую этот объект можно взять. В работе [25] авторы предлагают метод генерации синтетических обучающих данных для обучения робота. Сначала вручную делается несколько кадров объектов с разных сторон, указывается масштаб каждой из сторон, и указывается, какая сторона пригодна для захвата роботом. Затем из входных изображений система автоматически извлекает стороны объектов, выполняется масштабирование и построение виртуального объекта. Затем из этих виртуальных объектов создается куча, и на финальной стадии модуль генератора копирует и помещает созданные виртуальные объекты на то же изображение в случайном положении и со случайным вращением, создавая синтетическую кучу. Результаты показывают, что модель, обученная таким образом, не является конкурентом лучшим детекторам объектов, обученным на больших массивах реальных изображений, но хорошо подходит для конкретной задачи обнаружения объектов, которые можно поднимать в кучах.

Не менее интересный вариант использования рандомизации представлен в работе [26], где авторы предлагают новый метод создания чисто синтетических обучающих данных для обнаружения объектов. Авторы используют большой набор 3D-моделей фона и выполняют их плотную визуализацию с использованием полной рандомизации домена. В результате получаются фоновые изображения с реалистичными формами и текстурой, поверх которых визуализируются интересующие объекты. Во время обучения процесс генерации данных следует стратегии обучения, гарантирующей, что все модели переднего плана будут предварительно отправлены в нейросеть одинаково во всех возможных позах и условиях с возрастающей сложностью. В результате полностью контролируется основная статистика и создаются оптимальные обучающие выборки на каждом этапе обучения. Используя набор из 64 объектов розничной торговли, демонстрируется, что данный подход позволяет обучать детекторы, которые превосходят модели, обученные на реальных данных, на сложном наборе оценочных данных.

Представленные работы показывают, что методы, основанные на принципах рандомизации, могут быть одинаково эффективно использованы как при генерации табличных обучающих данных, так и для генерации данных для обучения систем распознавания изображений.

Генеративные методы создания синтетических данных

В основе одной из самых современных и продвинутых технологий генерации синтетических данных лежит использование генеративных состязательных сетей (GAN). Эта технология используется, например, в области безопасности и компьютерной безопасности, в области социальных наук, в медицине и компьютерном зрении. Далее будут представлены публикации, описывающие эти методы.

В области компьютерной безопасности исследователи все чаще используют методы машинного обучения (ML) для разработки и внедрения систем обнаружения вторжений (IDS) для компьютерных сетей. Многие из этих исследователей использовали наборы данных, собранные различными организациями, для обучения ML-моделей для прогнозирования вторжений. Во многих наборах данных, используемых в таких системах, данные несбалансированны, вследствие чего модели прогнозирования могут давать неудовлетворительные результаты классификации, что влияет на точность прогнозирования вторжений. В работе [27] авторы используют метод генерации синтетических данных под названием Conditional Generative Adversarial Network (CTGAN) для балансировки данных и изучения их влияния на различные ML-классификаторы. На основе обширных экспериментов с использованием широко используемого набора данных NSL-KDD[10] авторы обнаружили, что обучение ML-моделей на наборе данных, сбалансированном с помощью синтетических образцов, сгенерированных CTGAN, увеличивает точность предсказания на 8 % по сравнению с обучением тех же ML-моделей на несбалансированных данных.

Основной проблемой, с которой сталкиваются организации, пытающиеся предотвратить «отмывание» денег через азартные игры, также является отсутствие высококачественных данных. Поскольку при обнаружении мошенничества возникает серьезная проблема дисбаланса классов, в работе [28] авторы предлагают новую систему на основе генеративных состязательных сетей (GAN) для генерации синтетических данных с целью обучения классификатора под наблюдением. Описанная авторами система Synthetic Data Generation GAN (SDG-GAN) позволяет улучшить классификационные характеристики эталонных наборов данных и реального набора данных по мошенничеству в сфере азартных игр.

В работе [29] авторы рассматривают потенциальное применение GAN для создания синтетических микроданных переписи населения. Они используют ряд показателей полезности и метрику риска раскрытия информации (целевая вероятность правильной атрибуции) для сравнения данных, полученных с помощью табличных GAN, с данными, полученными с помощью ортодоксальных методов синтеза данных.

В статье [30] представлены методы генерации синтетических медицинских изображений с использованием генеративных сетей глубокого обучения GAN, показано, что созданные медицинские изображения могут быть использованы для пополнения синтетических данных и улучшения производительности сверточных сетей (convolutional neural networks – CNN) для классификации медицинских изображений. Предложенный метод продемонстрирован на ограниченном наборе данных компьютерной томографии (КТ) изображений 182 поражений печени (53 кисты, 64 метастаза и 65 гемангиом). Сначала используется архитектура GAN для синтеза высококачественных ROI поражений печени. За-

[10] NSL-KDD – набор данных, который может быть использован в качестве эффективного эталонного набора данных, чтобы помочь исследователям сравнить различные методы обнаружения вторжений [Электронный ресурс]. – URL: https://www.unb.ca/cic/datasets/nsl.html (дата обращения: 29.03.2023).

тем представляется новая схема классификации поражений печени с использованием CNN. Далее CNN обучается с использованием классического дополнения данных и синтетического дополнения данных. При добавлении синтетических данных эффективность результатов увеличилась с 78,6 до 85,7 %.

В работе [31] авторы сосредоточились на генерации синтетических мульти-последовательных магнитно-резонансных (МР) изображений головного мозга с помощью генеративных сетей (GAN). Предложенный подход к созданию реалистичных медицинских изображений показывает, что GAN могут генерировать 128×128 МР-изображения мозга, избегая артефактов. На рис. 3 и 4 представлены реальные и синтетические изображения соответственно.

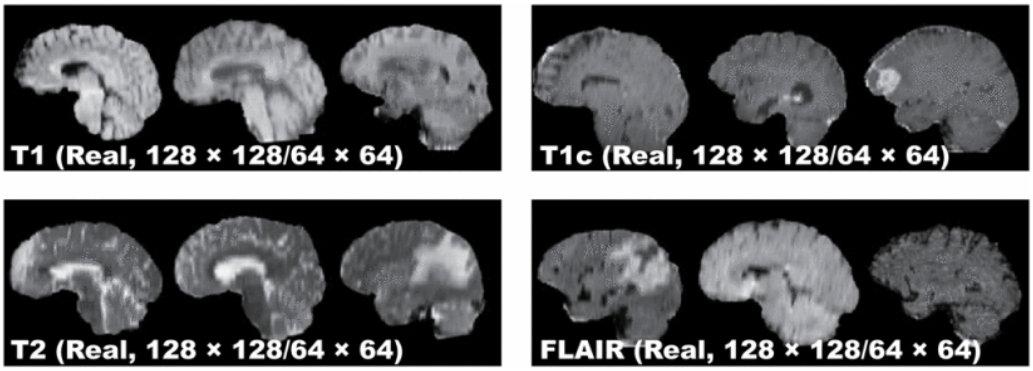


Рис. 3. Реальные МРТ-изображения головного мозга [30]

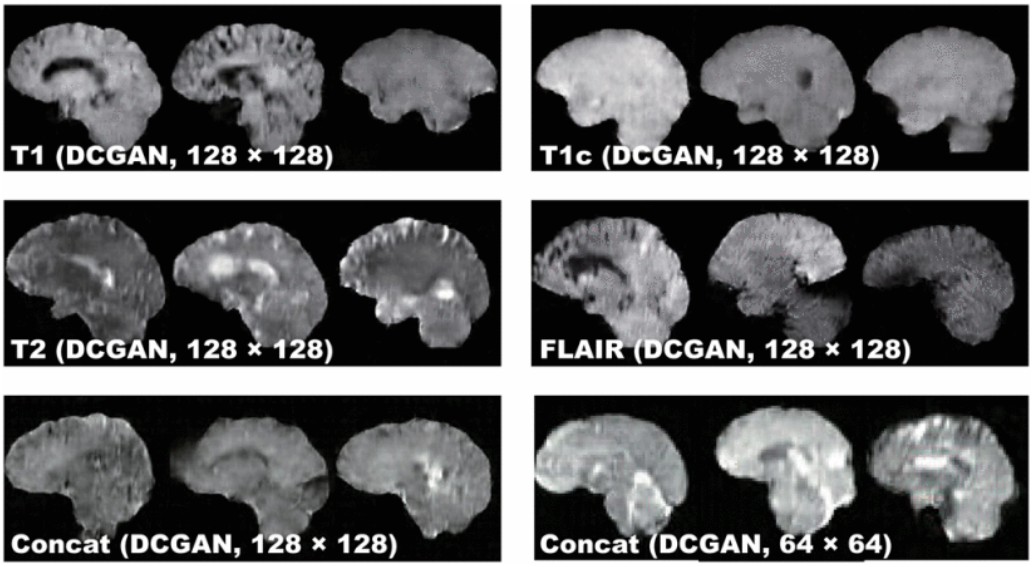


Рис. 4. Синтетические МРТ-изображения головного мозга [30]

В ходе предварительной проверки даже врач-эксперт не смог точно отличить синтетические изображения от реальных образцов в визуальном тесте Тьюринга.

Обмен наборами данных медицинской визуализации между учреждениями и даже внутри одного учреждения ограничен различными нормативно-правовыми барьерами. В результате медицинские исследовательские проекты, требующие больших наборов данных, значительно страдают. В последние годы в машинном обучении произошла революция с появлением подходов на основе глубоких нейронных сетей, что делает ограничения, связанные с данными, еще более серьезной проблемой, поскольку эти новые методы обычно требуют ог-

ромных наборов данных визуализации. В работе [32] представлены ограниченные ансамбли генеративных состязательных сетей (cGANe) для решения этой проблемы путем изменения представления данных визуализации, сохраняя при этом важную информацию, что позволяет воспроизводить аналогичные результаты исследований в других местах с помощью данных, доступных для обмена. Также описана структура, представляющая генерацию cGANe. Данный подход проверен на генерации синтетических трехмерных данных о метастатических областях мозга из T1-взвешенных МРТ-изображений с контрастным усилением. При 90%-ной чувствительности обнаружения метастазов головного мозга (МГМ) ранее представленный алгоритм обнаружения в среднем давал 9,12 ложноположительного обнаружения МГМ на пациента после обучения на исходных данных, в то время как после обучения на синтетических данных, сгенерированных cGANe, он давал 9,53 ложноположительных результата. Хотя применимость представленного подхода требует дальнейших валидационных исследований с использованием ряда типов данных медицинской визуализации, результаты показывают, что алгоритм обнаружения может достичь сопоставимой производительности при использовании синтетических данных, сгенерированных cGANe.

Сбор данных, ориентированный на человека, как правило, требует больших затрат и связан с вопросами конфиденциальности. В литературе предлагались различные решения для снижения этих затрат, например, сбор данных с помощью толпы или использование полусамостоятельных алгоритмов. Однако полусамостоятельные алгоритмы требуют источника немаркированных данных, а методы краудсорсинга требуют большого количества активных участников. Альтернативным методом пассивного сбора данных может быть использование уровня принимаемого сигнала (RSS) или информации о состоянии канала (CSI) в беспроводных сенсорных сетях для локализации пользователей в помещениях и на улице. В статье [33] представлен новый подход к снижению затрат на сбор обучающих данных за счет использования синтетических данных. Генеративные состязательные сети (GAN) используются для изучения распределения ограниченной выборки собранных данных и, следуя этому, для создания синтетических данных, которые могут быть применены для дополнения реальных собранных данных с целью повышения общей точности позиционирования. Результаты экспериментов на эталонном наборе данных показывают, что, применяя предложенный метод и используя комбинацию из 10 % собранных данных и 90 % синтетических данных, можно получить точность позиционирования, практически аналогичную той, которая была бы получена при использовании полного набора собранных данных. Это означает, что, используя синтетические данные, генерируемые GAN, можно использовать на 90 % меньше реальных данных, тем самым снижая затраты на сбор данных и достигая приемлемой точности.

В статье [34] представлена усовершенствованная схема генерации и адаптации синтетических изображений для обучения глубоких сверточных сетей для выполнения задачи обнаружения объектов в интеллектуальных торговых автоматах. Для решения проблемы, связанной с генерацией виртуальных изображений, похожих на сложные реальные сцены, и минимизацией избыточных обучающих данных, авторы рассматривают моделирование беспорядочных объектов, размещенных в виртуальной сцене, и использование широкоугольной камеры с искажениями, применяемой для захвата всей сцены в процессе генерации данных, а также постобработку сгенерированных изображений с помощью тщательно разработанной генеративной сети, чтобы сделать их более похожими на реальные изображения.

В статье [35], решая проблему доступности больших наборов данных, авторы предложили новые технологии генерации и дополнения синтетических данных для улучшения

обучения на спутниковых снимках с низким/нулевым количеством образцов. В дополнение к расширению подходов к генерации синтетических данных они предлагают иерархический подход к обнаружению для повышения полезности синтетических обучающих образцов. Авторы анализируют существующие методы создания синтетических изображений – 3D-модели и перенос нейронного стиля, а также вводят свою собственную рескиннинговую сеть GAN-Reskinner, обученную в неблагоприятных условиях для смешивания 3D-моделей. Чтобы проверить эффективность синтетических изображений, авторы оценивают полученные модели на реальных спутниковых снимках. Эксперименты показывают, что синтетические данные, разработанные с помощью данного подхода, часто повышают эффективность обнаружения, особенно в сочетании с реальными учебными изображениями.

Также интерес представляет работа [36], в которой исследуется метод обеспечения дифференциальной конфиденциальности генератора GAN. Полученная модель используется для создания синтетических данных, на которых можно обучать и тестировать алгоритмы, а также проводить испытания, не нарушая конфиденциальности исходного набора данных. Предложенный метод модифицирует систему Private Aggregation of Teacher Ensembles (PATE) и применяет ее к GAN. Модифицированная структура (PATE-GAN) позволяет жестко ограничить влияние любой отдельной выборки на модель, что приводит к жестким дифференциальным гарантиям конфиденциальности и, следовательно, к улучшению производительности по сравнению с моделями с такими же гарантиями. Эксперименты на различных наборах данных показывают, что относительная производительность модели на синтетических данных PATE-GAN стабильно превосходит производительность на оригинальных данных.

В статье [37] также показано, как генерировать дифференцированные частные синтетические данные с помощью GAN. Авторы объединяют три идеи: создание искусственных копий исходных данных, синтетические микроданные и байесовский GAN.

Дифференциальная конфиденциальность обеспечивает надежную гарантию конфиденциальности для защиты отдельных записей данных, содержащих конфиденциальную информацию. В работе [38] предлагается MC-GEN – метод генерации синтетических данных с сохранением конфиденциальности при гарантии дифференциальной конфиденциальности для задач множественной классификации. MC-GEN строит дифференциально частные генеративные модели на многоуровневых кластеризованных данных для генерации синтетических наборов данных и одновременно уменьшает шум, вносимый дифференциальной конфиденциальностью, для повышения полезности. В ходе экспериментальной оценки авторы оценили влияние параметров MC-GEN и сравнили MC-GEN с тремя существующими методами. Результаты сравнения показали, что MC-GEN может достичь значительной эффективности при определенных гарантиях конфиденциальности на нескольких задачах классификации.

В работе [39] представлена новая система совместного генеративного моделирования (CGM), которая стимулирует сотрудничество между сторонами, преследующими собственные интересы, для предоставления данных в пул для обучения генеративной модели (например, GAN), из которого берутся синтетические данные и распределяются между сторонами в качестве вознаграждения, соразмерного их вкладу. Распределение синтетических данных в качестве вознаграждения (вместо обученных моделей или денег) обеспечивает преимущества, связанные с задачами и моделями, для задач потокового обучения и с меньшей вероятностью нарушает правила конфиденциальности данных. Для реализации этой схемы авторы сначала предлагают функцию оценки данных с использованием максимального среднего расхождения (MMD), которая оценивает данные на основе их количества и качества с точки зрения их близости к истинному распределению данных, и предоставляют

теоретические результаты, определяющие выбор ядра в предложенной функции оценки данных на основе MMD. Затем они формулируют схему вознаграждения как линейную оптимизационную задачу, решение которой гарантирует определенные стимулы, такие как справедливость в рамках CGM. Авторы разрабатывают алгоритм взвешенной выборки для генерации синтетических данных, которые будут распределены между сторонами в качестве вознаграждения таким образом, чтобы стоимость их данных и синтетических данных в совокупности соответствовала назначенной им стоимости вознаграждения по схеме вознаграждения. Используя смоделированные и реальные наборы данных, они эмпирически доказывают, что вознаграждение сторон за синтетические данные соизмеримо с их вкладом.

Авторы статьи [40] изучают частную генерацию синтетических данных для выпуска запросов, где целью является построение санированной версии конфиденциального набора данных с учетом дифференциальной конфиденциальности, которая приблизительно сохраняет ответы на большую коллекцию статистических запросов. Представлена алгоритмическая схема, которая объединяет длинный ряд итерационных алгоритмов в литературе. В рамках этой структуры предлагаются два новых метода. Первый метод, Private Entropy Projection (PEP), можно рассматривать как усовершенствованный вариант MWEM, который адаптивно использует прошлые измерения запроса для повышения точности. Второй метод, генеративные сети с экспоненциальным механизмом (GEM), – позволяет обойти узкие места в вычислениях таких алгоритмов, как MWEM и PEP, путем оптимизации по генеративным моделям, параметризованным нейронными сетями, которые охватывают богатое семейство распределений и позволяют быстро оптимизировать на основе градиента. В статье демонстрируется, что PEP и GEM эмпирически превосходят существующие алгоритмы. Более того, авторы показывают, что GEM прекрасно учитывает предварительную информацию из открытых данных, преодолевая ограничения PMWPub, существующего современного метода, который также использует открытые данные.

Пример использования GAN для повышения качества съемки смартфонов при низких уровнях освещенности представлен в работе [41], в которой авторы решают проблему сбора данных для RAW-RGB видео при слабом освещении, предлагая механизм синтеза данных, названный SIDGAN, который может генерировать большое количество динамических обучающих пар видео. SIDGAN сопоставляет видео, найденные «в природе» (например, интернет-видео), с областью низкой освещенности (короткая или длинная экспозиция). Синтетическое генерирование динамических видеоданных позволяет недавно предложенной современной модели RAW-to-RGB достичь более высокого качества изображения (улучшение цвета, уменьшение артефактов) и улучшить временную согласованность, по сравнению с той же моделью, обученной только на статических реальных видеоданных.

Глубокие генеративные модели, такие как вариационные автоэнкодеры (VAE), являются популярным подходом для создания синтетических наборов данных на основе исходных данных. Несмотря на успех VAE, существуют ограничения, когда речь идет о бимодальных и перекошенных краевых распределениях. Они отклоняются от унимодальных симметричных распределений, которые поощряются предположением о нормальности, обычно используемым для латентных представлений в VAE. Хотя существуют расширения, предполагающие другие распределения для латентного пространства, это, как правило, не повышает гибкость для данных с большим количеством различных распределений. Поэтому авторы [42] предлагают новый метод – вариационные автоэнкодеры с предварительным преобразованием (PTVAEs) – для решения проблемы бимодальных и перекошенных данных, используя предварительные преобразования на уровне исходных переменных. Для приближения данных к

нормальному распределению используются два типа преобразований путем отдельной оптимизации параметров для каждой переменной в наборе данных. Они сравнивают производительность этого метода с другими современными методами генерации синтетических данных. В дополнение к визуальному сравнению авторы используют измерение полезности для количественной оценки. Результаты показывают, что подход PTVAE может превзойти другие методы при генерации бимодальных и перекошенных данных. Более того, простота подхода делает его пригодным для использования в сочетании с другими расширениями VAE.

Модели машинного обучения подвергаются критике за то, что они отражают погрешности в обучающих данных. Вместо того чтобы решать эту проблему путем введения корректных алгоритмов обучения напрямую, авторы [43] сосредоточились на создании корректных синтетических данных, чтобы любое последующее обучение было корректным. Генерирование правдивых синтетических данных из необъективных данных, оставаясь корректным по отношению к основному процессу генерирования данных (DGP), является нетривиальной задачей. В своей статье они представляют DECAF: генератор объективных синтетических данных на основе GAN для табличных данных. В DECAF они встраивают DGP в явном виде как структурную причинную модель во входные слои генератора, позволяя каждой переменной быть реконструированной в зависимости от ее причинных родителей. Эта процедура позволяет проводить дебайесинг во время вывода, когда смещенные грани могут быть стратегически удалены для удовлетворения заданных пользователем требований правдивости. Система DECAF универсальна и совместима с несколькими популярными определениями правдивости. Эксперименты показали, что DECAF успешно удаляет нежелательные смещения и – в отличие от существующих методов – способен генерировать высококачественные синтетические данные. Более того, авторы предоставляют теоретические гарантии сходимости генератора и правдивости последующих моделей.

Как видим, несмотря на различие предметных областей применения и разнообразие алгоритмов, все представленные методы содержат в своей основе генеративные состязательные сети (GAN). Следует отметить, что применение GAN для генерации синтетических данных в настоящее время является, пожалуй, самой распространенной практикой. В табл. 1 представлена сводная информация о методах, их особенностях и основных сферах применения.

Таблица 1

Основные свойства методов, основанных на генеративных состязательных сетях

Метод	Базовая архитектура	Способ использования	Тип данных	Особенность метода
4.1. GAN	GAN	Дополнение к реальным данным	Численные данные	Способность генерировать множество примеров, близких к оригиналу
4.2. SDG-GAN	cGAN	Дополнение, балансировка	Численные данные	Управление процессом генерации
4.3. CTGAN	cGAN	Дополнение, балансировка	Табличные данные, смешанные типы данных	Управление процессом генерации
4.4. DECAF	cGAN	Дополнение, балансировка	Табличные данные	Управление генерацией на основе причинно-следственных связей в данных
4.5. DCGAN	DCGAN	Дополнение	Изображения	Способность генерировать множество похожих изображений
4.7. SIDGAN	CycleGAN	Дополнение	Изображения	Создание парных образцов изображений
4.8. PATE-GAN	GAN	Создание, дополнение	Любой тип данных	Обеспечение дифференциальной приватности данных

Математические методы

Однако и математические методы также находят применение для генерации синтетических данных. Например, для построения синтетических версий небольшого по размеру, но реального набора данных, в работе [44] авторы предлагают парадигму множественной интерполяции. Анализ синтетических данных, полученных с помощью иерархических байесовских моделей, предложенных авторами, демонстрирует большую устойчивость к модели интерполяции, чем анализ синтетических данных, основанных на неиерархических версиях тех же моделей интерполяции.

Комбинация дифференциальной конфиденциальности и синтетических данных была представлена в качестве наилучшего решения в работе [45], где предлагается бесшумный метод построения дифференциально приватных синтетических данных. Авторы делают это с помощью механизма, называемого privatesampling. Используя булев куб в качестве эталонной модели данных, они вывели явные границы точности и конфиденциальности построенных синтетических данных. Ключевыми математическими инструментами являются гиперконтрактивность, двойственность и эмпирические процессы. Основным компонентом предложенного механизма частной выборки является строгий метод «граничной коррекции», который обладает замечательным свойством: перевзвешивание важности может быть использовано для точного соответствия граничных значений выборки граничным значениям популяции.

В работе [46] авторы сосредоточились на NP-трудной задаче разработки метода генерации синтетических данных, который был бы вычислительно эффективным, обеспечивал доказанные гарантии конфиденциальности и строго оценивал полезность данных. Авторы предлагают математические методы, которые позволяют им вывести конструктивные, приблизительно оптимальные решения сложных прикладных задач, касающихся микроагрегации, конфиденциальности и синтетических данных.

Несмотря на то, что GAN достигли значительных результатов в генерации синтетических данных, они часто сложны для интерпретации. Кроме того, методы на основе GAN могут пострадать при использовании смешанных реальных и категориальных переменных. Более того, дизайн функции потерь (потери дискриминатора) сам по себе специфичен для конкретной задачи, т.е. генеративная модель может оказаться бесполезной для задач, для которых она не была специально обучена. В работе [47] авторы предлагают использовать вероятностную модель в качестве генератора синтетических данных. По утверждению авторов, обучение вероятностной модели для генерации синтетических данных похоже на оценку плотности данных. Основываясь на теории копулы, они разделяют задачу оценки плотности на две части: оценка одномерных краевых значений и оценка плотности мультивариативной копулы над одномерными краевыми значениями. Авторы используют нормализующие потоки для изучения как копулятивной плотности, так и одномерных краевых значений и тестируют свой метод на симулированных и реальных наборах данных с точки зрения оценки плотности, а также способности генерировать высокоточные синтетические данные.

В статье [48] исследуется конкретная задача создания синтетических данных, называемая понижением масштаба. Это процедура вывода информации высокого разрешения, которую трудно собрать, из множества источников низкого разрешения, которые собрать гораздо легче. Авторами предлагается многоступенчатая структура под названием SYNC (Synthetic Data Generation via Gaussian Copula). Для заданных наборов данных низкого разрешения центральная идея SYNC состоит в том, чтобы подогнать модели гауссовой копулы к каждому из наборов данных низкого разрешения, чтобы правильно отразить зависимости и

граничные распределения, а затем сделать выборку из подогнанных моделей для получения желаемых поднаборов высокого разрешения. Затем прогнозные модели используются для объединения отобранных подмножеств в одно, и, наконец, отобранные наборы данных масштабируются в соответствии с предельными ограничениями низкого разрешения.

Авторы статьи [49] предлагают сквозной подход для создания синтетических данных вопросов и ответов (question/answering – QA). Их модель состоит из одной сети кодеров – декодеров на основе трансформаций, которая обучена генерировать как ответы, так и вопросы. В двух словах: на кодер они подают отрывок и просят декодер генерировать вопрос и ответ токен за токеном. Вероятность, полученная в процессе генерации, используется как показатель фильтрации, что позволяет избежать необходимости в отдельной модели фильтрации. Такой генератор обучается путем точной настройки предварительно обученной языковой модели (language models – LM) с помощью оценки максимального правдоподобия. Результаты экспериментов свидетельствуют о значительном улучшении адаптации моделей QA к домену по сравнению с современными методами.

Авторы [50] изучают сложность выборки частных генераторов синтетических данных для неограниченного класса статистических запросов и показывают, что любой класс, который является частным правильным PAC обучаемым[11], допускает частный генератор синтетических данных (возможно, неэффективный). Предыдущие работы по синтетическим генераторам данных были посвящены случаю, когда класс запросов D конечен, и были получены границы сложности выборки, которые масштабируются логарифмически с размером |D|. Авторы строят частный генератор синтетических данных, сложность выборки которого не зависит от размера области, и заменяют конечность предположением, что D является частным PAC-обучаемым.

Очевидно, что математические методы применяются там, где, по мнению исследователей, генеративные и другие методы неприменимы. В основном они используются для генерации текстовых данных или табличных данных смешанного типа.

Генерация синтетических данных с помощью симуляторов, CAD и виртуальных миров

Бурное развитие систем моделирования, симуляции и виртуальной реальности привело к тому, что эти технологии стали все чаще использоваться для генерации синтетических данных. В частности, виртуальные миры используются как среда, пригодная для генерации неограниченного количества синтетических данных, которые затем переносятся в реальный мир. И наоборот, реальные данные упрощаются и переносятся в виртуальный мир, где встречаются с моделями, обученными на синтетических данных из этих виртуальных миров. Наряду с технологиями виртуальных миров синтетические данные часто генерируются с помощью различных симуляторов и систем проектирования CAD.

Интересный подход к созданию синтетических данных из свободно доступных 3D CAD-моделей предлагается в работе [51]. Авторы утверждают, что, используя этот подход, можно легко генерировать бесконечное количество обучающих изображений практически для любого объекта. Авторы исследуют инвариантность сверточных сетей CNN к различным внутриклассовым вариациям, моделируя различные условия рендеринга, и получают удиви-

[11] Probably Approximately Correct learning (PAC learning) – схема машинного обучения, использующая понятия асимптотической достоверности и вычислительной сложности. Предложена в 1984 г. Лесли Вэлиантом.

тельные результаты. Основываясь на этих результатах, предлагается оптимальная стратегия генерации синтетических данных для обучения детекторов объектов на основе CAD-моделей.

Пытаясь решить проблему разрыва между синтетическими и реальными изображениями, возникающую в процессе сбора крупномасштабных данных изображений, авторы статьи [52] предлагают метод, который переносит обучение определению положения объекта из среды моделирования в реальный мир. Этот метод использует ограниченный набор данных реальных изображений и большой набор данных синтетических изображений с помощью вариационных автокодеров.

Обучение глубоких нейронных сетей для понимания изображений требует большого количества визуальных данных, специфичных для конкретной области. Хотя сбор таких данных от реальных роботов возможен, такой подход ограничивает масштабируемость, поскольку для обучения политик обычно требуются тысячи испытаний. В работе [53] предпринята попытка изучить политики манипулирования в симулированной среде. Симуляторы обеспечивают масштабируемость и предоставляют доступ к базовому состоянию окружающего мира во время обучения. Политики, выученные в симуляторах, однако не очень хорошо переносятся на реальные сцены, учитывая разницу между реальными и синтетическими данными. Авторы дополняют синтетические изображения последовательностями случайных преобразований. Основной их вклад заключается в оптимизации стратегии дополнения для переноса на реальные сцены и возможности обучения политике, не зависящей от области.

Действительно, симуляция все чаще используется для создания больших наборов данных с маркировкой во многих задачах машинного обучения. Последние методы были сосредоточены на настройке параметров симулятора с целью максимального повышения точности на задаче валидации, обычно полагаясь на градиентные оценки, подобные REINFORCE. Однако эти подходы очень дороги, поскольку они рассматривают всю линию генерации данных, обучения модели и валидации как «черный ящик» и требуют множества дорогостоящих оценок валидности на каждой итерации. В работе [54] авторы предлагают эффективную альтернативу для оптимальной генерации синтетических данных, основанную на новой дифференцируемой аппроксимации цели. Это позволяет им оптимизировать симулятор, который требует только одной оценки валидности на каждой итерации с небольшими накладными расходами. Они демонстрируют на современном фотореалистичном рендере, что предложенный метод до 50 раз быстрее находит оптимальное распределение данных, генерируя при этом в 30 раз меньше обучающих данных и обеспечивая на реальных тестовых наборах данных точность на 8,7 % выше, чем предыдущие методы.

Возможность использования видеоигр и симуляторов для создания крупномасштабных наборов данных семантической сегментации в реальном мире исследуется в работе [55]. Крупномасштабные синтетические наборы данных, как правило, используются как дополнения реальных наборов данных для обучения глубоких нейронных сетей. Другое применение синтетических наборов данных – это возможность проведения контролируемых и повторяемых экспериментов, благодаря возможности манипулировать содержанием и рендерингом синтезированных изображений. С этой целью авторы предложили метод генерации произвольно большого набора данных семантической сегментации, отражающего реальные особенности, при минимизации необходимых затрат и человеко-часов. Метод демонстрируется на примере создания ProcSy[12], синтетического набора данных для семантической сегментации,

[12] Официальная страница сайта, где хранится синтетический набор данных для семантической сегментации, который смоделирован на основе реальной городской среды и имеет ряд переменных

который смоделирован на основе реальной городской среды и имеет ряд переменных факторов влияния, таких как погода и освещение. В результате экспериментов было показано, что включение всего 3 % дождливых изображений в набор для обучения улучшает индекс оценки семантической сегментации (mIoU) сети на дождливых изображениях примерно на 10 %.

Точная локализация камеры является важной частью систем слежения. Однако на результаты локализации сильно влияет освещенность. Включение данных, собранных при различных условиях освещения, может повысить устойчивость алгоритма локализации к изменению освещения. Однако это очень утомительно и требует много времени. Используя синтетические изображения, можно легко накопить большое количество изображений при различном освещении и погодных условиях. Несмотря на постоянное совершенствование вычислительной мощности и алгоритмов рендеринга, синтетические изображения не полностью соответствуют реальным изображениям одной и той же сцены, т.е. существует разрыв между реальными и синтетическими изображениями, что также влияет на точность локализации камеры. Чтобы уменьшить влияние этого разрыва, в работе [56] авторы вводят преобразование реальных изображений в синтетические (REal-to-Synthetic Transform – REST). REST – это сеть, подобная автоэнкодеру, которая преобразует реальные признаки в их синтетический аналог. Затем преобразованные признаки могут быть сопоставлены с накопленной базой данных для надежной локализации камеры. Представленные результаты показывают, что REST повышает точность сопоставления примерно на 30 %.

Новый подход к синтезу изображений, которые эффективны для обучения детекторов объектов, предлагается авторами [57]. Начиная с небольшого набора реальных изображений, их алгоритм оценивает параметры рендеринга, необходимые для синтеза аналогичных изображений с учетом грубой 3D-модели целевого объекта. Эти параметры затем могут быть повторно использованы для создания неограниченного количества обучающих изображений интересующего объекта в произвольных 3D-позах, которые затем могут быть использованы для повышения эффективности классификации.

Центральным компонентом многообъектного слежения (Multi-object Tracking – MOT) является ассоциация, направленная на установление связи между с одинаковыми граничными полями в видеопоследовательности. Для обучения модулей ассоциации, например, параметрических сетей, обычно используются реальные видеоданные. Однако аннотирование следов человека в последовательных видеокадрах стоит дорого, а такие реальные данные в силу своей негибкости предоставляют ограниченные возможности для оценки эффективности системы при изменении сценариев слежения. В статье [58] авторы изучают, могут ли 3D-синтетические данные заменить реальное видео для обучения ассоциации. В частности, они представляют крупномасштабный механизм синтетических данных под названием MOTX[13], в котором характеристики движения камер и объектов вручную настраиваются таким образом, чтобы быть похожими на те, которые присутствуют в реальных наборах данных. Они показывают, что, по сравнению с реальными данными, знания об ассоциациях, полученные из синтетических данных, могут достичь очень похожей производительности на реальных тестовых наборах без использования методов адаптации домена. Это интригующее наблюдение объясняется двумя

факторов влияния, таких как погода и освещение [Электронный ресурс]. – URL: https://uwaterloo.ca/waterloo-intelligent-systems-engineering-lab/procsy (дата обращения: 05.07.2023).

[13] Упрощенная абстракция схемы потока состояний для децентрализованного управления состояниями мультипроцессных или мультиизолированных модулей, обеспечивающая последовательную абстракцию межмодульной связи [Электронный ресурс]. – URL: https://github.com/zeusCore/motx (дата обращения: 29.03.2023).

факторами. Прежде всего, 3D-движки могут хорошо моделировать факторы движения, такие как движение камеры, вид камеры и движение объекта, так что смоделированные видео могут обеспечить модули ассоциаций эффективными характеристиками движения. Кроме того, экспериментальные результаты показывают, что разрыв в области внешнего вида практически не мешает обучению ассоциативным знаниям.

В работе [59] представлен новый метод с открытым исходным кодом для автоматического извлечения облаков точек обнаружения света и дальности с аннотациями на уровне данных из симулятора. Виртуальный датчик может быть настроен для имитации различных реальных устройств, от двухмерных лазерных сканеров до трехмерных датчиков реального времени. Проведенные эксперименты показали, что использование дополнительных синтетических данных для обучения позволяет: 1) добиться заметного повышения точности; 2) сократить количество вручную маркированных реальных данных; 3) улучшить обобщение по всем наборам данных.

Поскольку аннотирование данных для точного подсчета людей в плотной толпе нереально, авторы [60] предлагают метод, который создает управляемые симуляции людей, движущихся в реальной среде; люди моделируются как синтетические гуманоиды с реалистичной внешностью, в то время как фоном является реальное изображение сцены. Управляя освещением, динамикой людей и плотностью сцены, имеется возможность генерировать практически бесконечное количество симуляций и очень большие аннотированные наборы данных. Таким образом, за счет использования симулированных наборов данных производительность обычной глубокой сети, используемой для задачи анализа толпы, может быть улучшена.

В работе [61] представлена новая методология генерации синтетических данных для обучения глубокой нейронной сети (Deep Neural Networks – DNN) для оценки карт глубины непосредственно по стереоизображениям подводных сцен. Предложенный метод проецирует реальные подводные изображения на ландшафты с произвольной высотой в рамках 3D-рендеринга. Эта процедура предоставляет пару синтетических стереоизображений и соответствующую карту глубины сцены, которые используются для обучения DNN оценки диспаритета. Благодаря этому процессу авторы учатся сопоставлять пространство подводных объектов с помощью контролируемого обучения без необходимости получения обширных реальных карт глубины подводного пространства для получения достоверных данных. В результате авторы демонстрируют повышенную точность реконструкции по сравнению с традиционными методами сопоставления признаков компьютерного зрения и современными DNN, обученными на синтетических наземных данных.

Поскольку симуляторы, системы проектирования CAD и платформы виртуальной реальности предназначены для обработки 3D-изображений, методы на их основе главным образом применяются для дополнения наборов реальных обучающих данных, используемых для разработки систем распознавания образов, детекторов положения тел в пространстве, систем слежения, систем распознавания людей в толпе и т.д. Потрясающие успехи в развитии платформ виртуальной реальности являются поистине неиссякаемым источником синтетических данных. В сочетании с развитием техники переноса обучения из виртуальности в реальность и наоборот существенно повышается возможность генерации огромного количества графических изображений.

Метод адаптации домена

Адаптация домена – это способность применять алгоритм, обученный в одном или нескольких «исходных доменах», к другому (но связанному) «целевому домену». Доменная адаптация является подкатегорией трансфертного обучения. При адаптации домена исходный

и целевой домены имеют одинаковое пространство признаков (но разные распределения); в отличие от этого, трансфертное обучение включает случаи, когда пространство признаков целевого домена отличается от пространства или пространств признаков источника [62].

В работе [63] авторы предлагают новую многозадачную глубокую сеть для обучения обобщенным высокоуровневым визуальным представлениям. Поскольку многозадачное обучение требует аннотаций для нескольких свойств одного и того же учебного примера, они используют синтетические изображения для обучения этой сети. Чтобы преодолеть разницу в области между реальными и синтетическими данными, предлагается использовать метод адаптации домена пространства признаков без надзора, основанный на состязательном обучении.

Радиочастотные датчики были недавно предложены в качестве нового метода для технологии обработки языка жестов. Они бесконтактны, эффективны в темноте и позволяют получить прямое измерение кинематики жестов благодаря использованию микродоплеровского эффекта. В работе [64] проводится углубленное сравнительное исследование кинематических свойств жестов, измеренных радиочастотными датчиками, как для свободно говорящих на языке ASL (American Sign Language)[14], так и для людей, имитирующих жесты. Кроме того, поскольку методы распознавания ASL с использованием глубокого обучения требуют большого количества обучающих данных, в данной работе исследуется влияние кинематики жестов и беглости речи на методы состязательного обучения для синтеза данных. Предлагаются два различных подхода к созданию синтетических обучающих данных:

1) состязательная адаптация домена для минимизации различий между данными имитации жестов и беглого жеста;

2) кинематически ограниченные генеративные состязательные сети для точного синтеза жестов. Результаты показывают, что кинематические расхождения между имитацией жестов и беглым жестом настолько значительны, что обучение на данных, непосредственно синтезированных из беглых жестов, дает более высокую производительность (93 % точности топ-5), чем адаптация имитации жестов (88 % точности топ-5) при классификации 100 знаков ASL.

Среди самых больших проблем, с которыми мы сталкиваемся при использовании нейронных сетей, обученных на данных волновых форм (т.е. сейсмических, электромагнитных или ультразвуковых), является их применение к реальным данным. Требование точных меток заставляет нас разрабатывать решения с использованием синтетических данных, где метки легко доступны. Однако синтетические данные часто не отражают реальность реального эксперимента, и в итоге мы получаем низкую производительность обученной нейронной сети на этапе вывода. В статье [65] описан новый подход к улучшению контролируемого обучения на синтетических данных с учетом особенностей реальных данных (адаптация к домену). В частности, для задач, в которых абсолютные значения вертикальной оси (время или глубина) входных данных не имеют решающего значения, например, классификация, или могут быть скорректированы впоследствии, например, построение модели скорости с использованием журнала скважин, авторы предлагают выпол-

[14] Американский язык жестов (ASL) является естественным языком, который служит преобладающим языком жестов в сообществах глухих в Соединенных Штатах Америки и большей части англоязычной Канады. ASL – это законченный и организованный визуальный язык, который выражается с помощью как ручных, так и немануальных функций. Помимо Северной Америки, диалекты ASL и креольские языки на основе ASL используются во многих странах по всему миру, включая большую часть Западной Африки и части Юго-Восточной Азии.

нить ряд линейных операций над входными данными, чтобы обучающие и прикладные данные имели схожие распределения. Это достигается путем применения двух операций над входными данными для модели NN:

1) кросс-корреляция входных данных (т.е. сейсмограммы, сейсмического изображения и т.д.) с фиксированной эталонной трассой из того же набора данных;

2) свертка полученных данных со средним значением (или случайной выборкой) автокоррелированных данных из другой области.

На этапе обучения входные данные берутся из синтетической области, а автокоррелированные данные – из реальной области, и случайные выборки из реальных данных берутся в каждый период обучения. На этапе вывода/применения входные данные берутся из области реального подмножества, а средние значения автокоррелированных участков – из области синтетического подмножества данных. Примеры применения пассивных сейсмических данных для определения местоположения источника микросейсмических событий и активных сейсмических данных для прогнозирования низких частот используются для демонстрации возможностей этого подхода в улучшении применимости обученных моделей к реальным данным.

В настоящее время большинство существующих моделей слепой оценки качества изображений (Blind Image Quality Assessment – BIQA) разработаны для синтетически искаженных изображений и часто плохо применимы к подлинным. Кроме того, они сильно зависят от человеческих оценок, сбор которых непомерно трудозатратен. В работе [66] предлагается метод BIQA; который на основе синтетически искаженных изображений и множества агентов учится оценивать перцептивное качество подлинно искаженных изображений, полученных в естественных условиях, не полагаясь на человеческие оценки. В частности, сначала авторы собирают большое количество пар изображений из синтетически искаженных изображений и используют набор моделей оценки качества изображения (Full-reference Image Quality Assessment – FR-IQA) для присвоения псевдобинарных меток каждой паре, указывающих в качестве контролирующего сигнала, какое изображение имеет более высокое качество. Затем они обучают модель BIQA на основе сверточной нейронной сети CNN для ранжирования качества восприятия, оптимизированного для согласованности с двоичными метками. Поскольку между синтетически и достоверно искаженными изображениями существует сдвиг домена, для решения этой проблемы вводится модуль адаптации домена без контроля (Unsupervised Domain Adaptation – UDA).

Методы на основе адаптации домена по сути являются попыткой применения моделей, обученных на одной предметной области, в других предметных областях с помощью различных техник. Как правило, эти методы используются для обработки графических данных.

Системы генерации синтетических данных

В настоящее время сформировалась целая индустрия по производству синтетических данных, что было обусловлено многочисленными исследованиями в области генерации синтетических данных, а также все возрастающими потребностями в огромном количестве данных для обучения широкого спектра нейросетевых моделей. Сегодня на рынке доступно большое количество систем генерации синтетических данных, начиная от частных систем, предлагаемых отдельными группами исследователей, и заканчивая большими коммерческими платформами для производства синтетических данных в самых разнообразных областях. Так, в работе [67] авторы предлагают программу Meta-Sim, предназначенную для автоматического

синтеза наборов размеченных данных. Программа обучается генеративной модели синтетических сцен и создает изображения и соответствующую достоверную информацию с помощью графического движка. Генератор наборов данных настраивается с помощью нейронной сети, которая учится изменять атрибуты графов сцен, полученных из вероятностных графем сцен, так, чтобы минимизировать разрыв в распределении между своими выводами и целевыми данными. Эксперименты показывают, что предложенный метод может значительно улучшить качество генерации контента по сравнению с графемой сцены, разработанной человеком.

В статье [68] описывается система GeneSIS-RT, которая позволяет генерировать реалистичные данные для обучения реальным задачам, используя только немаркированные изображения реального мира и симуляцию. Предложенная система GeneSIS-RT облегчает бремя сбора маркированных реальных изображений и является перспективным кандидатом для создания высококачественных синтетических данных, специфичных для конкретной области. Алгоритмы обучения, натренированные на данных, полученных с помощью GeneSIS-RT, делают высокоточные прогнозы и превосходят системы, натренированные только на сырых симулированных данных, а также превосходят системы, натренированные на реальных данных.

Решение проблемы оценки глубины с помощью монокулярных камер позволяет широко использовать камеры в качестве недорогих датчиков оценки глубины в таких приложениях, как автономное вождение и робототехника. Однако для обучения такой масштабируемой модели оценки глубины потребуется большое количество маркированных данных, сбор которых стоит дорого. Существуют два популярных подхода, которые не требуют аннотированных карт глубины: (1) использование размеченных синтетических и неразмеченных реальных данных во внешнем окружении для более точного прогнозирования глубины и (2) модели без контроля, которые используют геометрическую структуру в пространстве и времени в монокулярных видеокадрах. Авторы работы [69] представляют систему самообучения S^3Net, которая использует синтетические и реальные изображения для обучения с учетом геометрических, временных, а также семантических ограничений. Эта оригинальная консолидированная архитектура обеспечивает новый уровень в самоконтролируемой оценке глубины с использованием монокулярного видео. Авторы представляют уникальный способ обучения этой самоконтролируемой системы и достигают более чем 15 % улучшения, по сравнению с предыдущими синтетическими контролируемыми подходами, использующими адаптацию домена, и более чем 10 % улучшения, по сравнению с предыдущими самоконтролируемыми подходами, использующими геометрические ограничения из реальных данных.

В последние годы обнаружение людей и оценка их позы достигли больших успехов благодаря крупномасштабным базам размеченных данных. Однако эти наборы данных не содержат гарантий или анализа действий человека, его позы или разнообразия контекста. Кроме того, конфиденциальность, правовые и этические проблемы могут ограничить возможность сбора большего количества данных о людях. Появившейся альтернативой реальным данным, которая снимает некоторые из этих проблем, являются синтетические данные. Однако создание генераторов синтетических данных невероятно сложно и не позволяет исследователям изучить их полезность. Поэтому авторы [70] выпускают ориентированный на человека генератор синтетических данных PEOPLESANSPEOPLE[15], который содержит готовые для моде-

[15] PeopleSansPeople – генератор синтетических данных для компьютерного зрения, ориентированного на человека [Электронный ресурс]. – URL: https://unity-technologies.github.io/PeopleSansPeople/ (дата обращения: 29.03.2023).

лирования 3D-активы человека, параметризованную систему освещения и камеры, а также генерирует 2D и 3D bounding box, экземпляр и семантическую сегментацию, а также метки позы COCO[16]. Используя PEOPLESANSPEOPLE, провели эталонное обучение синтетических данных с помощью варианта R-CNN Detectron2 Keypoint[17] и обнаружили, что предварительное обучение сети на синтетических данных и тонкая настройка на целевых данных реального мира (передача нескольких снимков на ограниченные подмножества COCO-person train) привели к AP по ключевой точке 60,37 ± 0,48 (COCO test-dev2017[18]), превосходящим модели, обученные только на тех же реальных данных (AP по ключевой точке 55,80) и предварительно обученные на ImageNet (AP по ключевой точке 57,50). Этот свободно распространяемый генератор данных должен позволить провести широкий спектр исследований в развивающейся области обучения с переходом от симуляции к реальным данным в критической области компьютерного зрения, ориентированного на человека.

Многие специалисты по исследованию данных используют различные инструменты для помощи в процессе создания и маркировки данных. Однако многие из них по-прежнему требуют взаимодействия с пользователем на протяжении всего процесса. Кроме того, большинство из них ориентированы только на несколько сетевых структур. Исследователи, изучающие несколько фреймворков, должны искать дополнительные инструменты или писать скрипты преобразования. В работе [71] представлен автоматизированный инструмент[19] для генерации синтетических данных в произвольных сетевых форматах. Он использует Robot Operating System (ROS)[20] и Gazebo[21], которые являются распространенными инструментами в сообществе робототехники. Благодаря парадигмам ROS, он позволяет пользователю широко настраивать среду моделирования и процесс генерации данных. Кроме того, плагиноподобная структура позволяет разрабатывать произвольные устройства записи форматов данных без необходимости изменения основного кода. Используя этот инструмент, авторы смогли сгенерировать произвольно большой набор данных изображений для трех уникальных форматов обучения,

[16] COCO имеет несколько типов аннотаций: для обнаружения объектов, обнаружения ключевых точек, сегментации вещей, паноптической сегментации, плотности и подписей к изображениям. Аннотации хранятся с помощью JSON. Обратите внимание, что для доступа и работы со всеми аннотациями можно использовать COCO API, описанный на странице загрузки [Электронный ресурс]. – URL: https://cocodataset.org/#format-data (дата обращения: 29.03.2023).

[17] Официальная страница сайта с документацией на Detectron2 [Электронный ресурс]. – URL: https://detectron2.readthedocs.io/en/latest/tutorials/index.html (дата обращения: 29.03.2023).

[18] Ссылка на скачивание датасета [Электронный ресурс]. – URL: http://images.cocodataset.org/zips/test2017.zip (дата обращения: 30.03.2023).

[19] Пакет ROS предназначен для автоматической генерации маркированных данных для обучения нейронных сетей с помощью Gazebo. Этот пакет предоставляет средства для автоматической генерации маркированных данных для обучения с использованием Gazebo для генерации синтетических данных [Электронный ресурс]. – URL: https://github.com/Navy-RISE-Lab/nn_data_collection (дата обращения: 30.03.2023).

[20] Robot Operating System (ROS) – набор программных библиотек и инструментов с открытым исходным кодом, которые помогают создавать приложения для роботов, включая драйверы, а также самые современные алгоритмы и мощные инструменты разработчика [Электронный ресурс]. – URL: https://www.ros.org/ (дата обращения: 30.03.2023).

[21] Gazebo – среда моделирования, которая предлагает новый подход к моделированию с полным набором библиотек для разработки и облачных сервисов для упрощения моделирования. Обеспечивает быструю итерацию новых физических конструкций в реалистичной среде с потоками датчиков высокой точности. Позволяет тестировать стратегии управления в условиях безопасности и использовать преимущества моделирования в тестах непрерывной интеграции [Электронный ресурс]. – URL: https://gazebosim.org/home (дата обращения: 29.03.2023).

затратив примерно 15 мин времени на настройку и переменное количество времени на выполнение, в зависимости от размера набора данных.

В 2019 г. был представлен инструмент генерации UnrealROX[22], позволяющий генерировать высокореалистичные данные с высоким разрешением и частотой кадров с помощью эффективного конвейера на базе Unreal Engine[23], передового движка для видеоигр. UnrealROX позволил исследователям роботизированного зрения генерировать реалистичные и визуально правдоподобные данные с полной достоверностью для решения широкого круга задач, таких как сегментация классов и экземпляров, обнаружение объектов, оценка глубины, визуальное грассирование и навигация. Тем не менее его рабочий процесс был очень привязан к генерации последовательностей изображений с бортовой камеры робота, что затрудняло генерацию данных для других целей. В работе [72] представлен UnrealROX+[24], улучшенная версия UnrealROX, в которой независимая и простая в использовании система сбора данных позволяет быстро разрабатывать и генерировать данные гораздо более гибким и настраиваемым способом. Более того, она упакована как плагин Unreal, что делает ее более удобной для использования с уже существующими проектами Unreal, а также включает новые функции, такие как генерация альбедо или API Python для взаимодействия с виртуальной средой из фреймворков Deep Learning.

В работе [73] предлагается новая среда для обучения, где входной областью является изображение карты, определенной на произведении двух множеств, одно из которых полностью определяет метки, а также представляется алгоритм, направленный на минимизацию биастермы путем использования возможности независимой выборки из каждого набора. Авторы применяют свой подход к задачам визуальной классификации, где он позволяет обучать классификаторы на наборах данных, состоящих полностью из одного синтетического примера каждого класса. На нескольких стандартных эталонах для классификации реальных изображений они достигают высокой производительности в контекстно-агностической постановке, с хорошей обобщенностью на реальные области, в то время как обучение непосредственно на реальных данных без использования этих методов дает классификаторы, хрупкие к возмущениям фона.

Несмотря на то, что трансферное обучение для GAN успешно улучшает производительность генерации в режимах с малым количеством примеров, предварительно обученная модель, использующая один эталонный набор данных, не обобщается на различные целевые наборы данных и может быть уязвима к рискам авторского права или конфиденциальности. Для решения обеих проблем в работе [74] предлагается эффективный и бес-

[22] UnrealROX – чрезвычайно фотореалистичная среда виртуальной реальности, созданная на движке Unreal Engine 4 и предназначенная для генерирования синтетических данных для различных задач роботизированного зрения. Эта среда виртуальной реальности позволяет исследователям роботизированного зрения генерировать реалистичные и визуально правдоподобные данные с полной достоверностью для решения широкого спектра задач, таких как семантическая сегментация классов и экземпляров, обнаружение объектов, оценка глубины, визуальный захват, навигация и др. [Электронный ресурс]. – URL: https://unrealrox.readthedocs.io/en/master/ (дата обращения: 30.03.2023).

[23] Unreal Engine (UE) – игровой движок для трехмерной компьютерной графики, разработанный компанией Epic Games и впервые продемонстрированный в 1998 г. в шутере от первого лица Unreal.

[24] UnrealROX+ – плагин для Unreal Engine 4, который позволяет легко получать разнообразные данные из виртуальной 3D-среды. Эти данные варьируются от данных изображений (RGB, карты глубины, карты нормалей, альбедо или маски сегментации) до 6D поз объектов, камер или скелетов. Репозиторий, где хранится программный код [Электронный ресурс]. – URL: https://github.com/3dperceptionlab/unrealrox-plus (дата обращения: 30.03.2023).

пристрастный синтезатор данных Primitives-PS. Авторы используют 1) общую статистику спектра частот, 2) элементарную форму (т.е. композицию изображений через элементарные формы) для представления информации о структуре; 3) существование освещенности в качестве предварительного условия. Поскольку синтезатор Primitives-PS учитывает только общие свойства естественных изображений, единственная модель, предварительно обученная на одном наборе данных, может быть последовательно перенесена на различные целевые наборы данных и даже превосходит предыдущие методы, предварительно обученные на естественных изображениях, по расстоянию приема Фреше. Обширный анализ демонстрирует эффективность предложенного синтезатора данных и дает представление о желательной природе предварительно обученной модели для переносимости GANs.

В процессе интеллектуального сегментирования продуктов питания на изображениях с помощью глубоких нейронных сетей для управления питанием сбор данных и маркировка для обучения сети являются очень важными, но трудоемкими задачами. Чтобы решить трудности сбора данных и аннотаций, в работе [75] предлагается метод сегментации продуктов питания, применимый к реальному миру через синтетические данные. Для выполнения сегментации пищи в робототехнических системах здравоохранения, таких как робот-манипулятор для помощи в приеме пищи, авторы генерируют синтетические данные с помощью открытого программного обеспечения для 3D-графики Blender, размещая несколько объектов на тарелке с едой и обучая Mask R-CNN для сегментации экземпляров. Кроме того, они создали систему сбора данных и проверили свою модель сегментации на реальных данных о еде. В результате на их наборе реальных данных модель[25], обученная только на синтетических данных, способна сегментировать не обученные экземпляры еды с 52,2 % маски AP@all, а после тонкой настройки производительность увеличилась на +6,4 % по сравнению с моделью, обученной с нуля. Кроме того, также подтверждается возможность и улучшение производительности на публичном наборе данных для справедливого анализа.

Пакет Synthpop[26] для R предоставляет инструменты, позволяющие хранителям данных создавать синтетические версии конфиденциальных микроданных, которые могут распространяться с меньшими ограничениями, чем оригиналы. Синтез может быть настроен таким образом, чтобы в синтетических данных обеспечить воспроизведение таких же взаимосвязей, как и в реальных данных. Для оценки этого аспекта был предложен ряд показателей, известных как полезность синтетических данных. В статье [76] показано, что все эти меры, включая те, которые рассчитываются на основе табуляций, могут быть получены на основе модели баллов склонности. В результате сравнения авторы показали, что все сравниваемые меры имеют высокую корреляцию, а некоторые даже идентичны. Метод, используемый для определения модели показателей склонности, более важен, чем выбор меры. Эти меры и методы включены в полезные модули пакета Synthpop, который включает методы визуализации результатов и, таким образом, обеспечивает немедленную обратную связь, позволяющую человеку, создающему синтетические данные, улучшить их качество. Утилитарные функции изна-

[25] Набор данных для сегментации объектов питания [Электронный ресурс]. – URL: https://github.com/gist-ailab/Food-Instance-Segmentation (дата обращения: 27.03.2023).

[26] Пакет synthpop для R позволяет пользователям создавать синтетические версии конфиденциальных данных индивидуального уровня для использования исследователями, заинтересованными в том, чтобы делать выводы о населении, которое представляют эти данные. Синтезированные данные могут быть опубликованы с меньшими ограничениями в отношении того, как они должны храниться, чем оригинальные данные [Электронный ресурс]. – URL: https://www.synthpop.org.uk/ (дата обращения: 30.03.2023).

чально были разработаны для использования на объектах синтетических данных classsynds, созданных функцией synthpop syn() или syn.strata(), но теперь они могут быть использованы для сравнения одного или нескольких синтезированных наборов данных с исходными записями, где записи представляют собой кадры данных R или списки кадров данных.

В работе [77] представлен пакет Unity Perception[27], который призван упростить и ускорить процесс создания синтетических наборов данных для задач компьютерного зрения, предлагая простой в использовании и хорошо настраиваемый набор инструментов. Этот пакет с открытым исходным кодом расширяет редактор Unity и компоненты движка для создания идеально аннотированных примеров для нескольких распространенных задач компьютерного зрения. Кроме того, он предлагает расширяемую структуру рандомизации, которая позволяет пользователю быстро создавать и настраивать рандомизированные параметры моделирования, чтобы внести вариативность в генерируемые наборы данных. Авторы предлагают обзор представляемых инструментов и их работы, а также демонстрируют ценность сгенерированных синтетических наборов данных путем обучения модели обнаружения 2D-объектов. Модель, обученная на преимущественно синтетических данных, превосходит модель, обученную только на реальных данных.

Чтобы обучить модели глубокого обучения для распознавания действий пожилых людей на основе зрения, обычно необходимы крупномасштабные наборы данных о действиях, полученных в различных условиях повседневной жизни. Однако большинство публичных наборов данных, используемых для распознавания действий человека, либо отличаются от наборов данных, используемых для распознавания действий пожилых людей, либо имеют ограниченный охват по многим аспектам, что затрудняет распознавание повседневной деятельности пожилых людей, если использовать только существующие наборы данных. В последнее время такие ограничения имеющихся наборов данных активно компенсируются путем генерирования синтетических данных из реалистичных сред моделирования и использования этих данных для обучения моделей глубокого обучения. В работе [78] на основе этих идей авторы разработали ElderSim[28] – платформу для моделирования действий, которая может генерировать синтетические данные о повседневной деятельности пожилых людей. Для 55 видов частых повседневных действий пожилых людей ElderSim генерирует реалистичные движения синтетических персонажей с различными настраиваемыми опциями генерирования данных и предоставляет различные выходные модальности, включая RGB-видео, двух- и трехмерные скелетные движения. Кроме того, на основе ElderSim авторы создали крупномасштабный синтетический набор данных о повседневной жизни пожилых людей KISTSynADL и используют эти данные в дополнение к реальным наборам данных для обучения трех современных моделей распознавания действий человека. В ходе экспериментов, проведенных по нескольким новым сценариям, предполагающим различные конфигурации реальных и синтетических наборов данных для обучения, наблюдается заметное улучшение производительности за счет дополнения синтетических данных.

Сведения, содержащиеся в данном разделе, могут быть полезны тем исследователям, которые нуждаются в дополнении или балансировке своих обучающих данных, но не имеют возможности разрабатывать собственные методы генерации синтетических данных.

[27] Инструментарий для обучения и проверки восприятия в режиме sim2real в Unity [Электронный ресурс]. – URL: https://github.com/Unity-Technologies/com.unity.perception (дата обращения: 30.03.2023).

[28] Официальная страница сайта проекта ElderSim [Электронный ресурс]. – URL: https://ai-4robot.github.io/ElderSim/ (дата обращения: 30.03.2023).

Коммерческие системы генерации данных

Много полезных сведений о производителях синтетических данных можно найти на сайте AI Multiple[29], которую создал Джем Дилмегани. Наиболее популярные случаи использования синтетических данных представлены на странице [79], а на страницах [80–82] представлено подробное руководство, в котором описаны различные области применения, а также выгоды от использования синтетических данных и методы их генерации. Данный сайт крайне полезен для исследователей, планирующих использовать синтетические данные, поскольку здесь представлены актуальные сведения по этой тематике. Постоянно обновляемый перечень наиболее популярных производителей синтетических данных представлен в табл. 2.

Таблица 2

Наиболее популярные производители синтетических данных

№ п/п	Наименование производителя	Дата создания	Количество сотрудников
1	BizDataX	2005	51-200
2	CA Technologies Datamaker	1976	10,001+
3	CVEDIA	2016	11-50
4	Deep Vision Data by Kinetic Vision	1985	51-200
5	Delphix Test Data Management	2008	501-1000
6	Genrocket	2012	11-50
7	Hazy	2017	11-50
8	Informatica Test Data Management Tool	1993	5,001-10,000
9	Mostly AI	2017	11-50
10	Neuromation	2016	11-50
11	Solix EDMS	2002	201-500
12	Supervisely	2017	2-10
13	TwentyBN	2015	11-50

Также представлен рейтинг[30] производителей, где наиболее популярным производителем является компания MOSTLY AI[31], которая предлагает ведущую, наиболее точную платформу синтетических данных, позволяющую предприятиям разблокировать, обмениваться, исправлять и моделировать данные. Благодаря достижениям в области искусственного интеллекта, синтетические данные MOSTLY AI выглядят и ощущаются так же, как реальные данные, и способны сохранять ценную детализированную информацию, гарантируя при этом их полную конфиденциальность. На втором месте располагается компания GenRocket[32], и следом за ней на третьей позиции представлена компания MDclone[33]. Всего в рейтинге представлены 19 производителей синтетических данных, что говорит о том, что индустрия по их производству уже оформлена как отдельный вид бизнеса.

[29] Официальная страница сайта [Электронный ресурс]. – URL: https://research.Aimultiple.com/synthetic-data/#synthetic-data-tools (дата обращения: 30.03.2023).

[30] Официальная страница сайта рейтинга наиболее популярных платформ генерации синтетических данных [Электронный ресурс]. – URL: https://aimultiple.com/synthetic-data-generator/1 (дата обращения: 30.03.2023).

[31] Официальная страница сайта MOSTLY AI [Электронный ресурс]. – URL: https://mostly.ai/synthetic-data-platform/ (дата обращения: 30.03.2023).

[32] Официальная страница сайта GenRocket [Электронный ресурс]. – URL: https://www.genrocket.com/ai-and-ml/ (дата обращения: 30.03.2023).

[33] Официальная страница сайта MDclone [Электронный ресурс]. – URL: https://www.mdclone.com/ (дата обращения: 30.03.2023).

Заключение

На основании представленного обзора работ в области генерации синтетических обучающих данных можно сделать вывод, что для генерации графических данных чаще всего используются методы на основе симуляторов и виртуальной реальности, GAN и адаптации домена. В то же время для обработки различных видов табличных данных, как правило, используются общие подходы, математические методы и методы на основе рандомзации.

Целью данной работы было ознакомить читателей с публикациями о синтетических данных, представить, что такое синтетические данные, для чего они используются, показать, в каких областях науки и техники они применяются, какие методы генерации синтетических данных существуют, с помощью каких инструментов их можно создавать.

Следует отметить, что многие исследователи так или иначе создают или используют синтетические данные в своих проектах. О возможности использования синтетических данных для обучения нейронных сетей, было доложено в 2008 г. на Международной научно-практической конференции в г. Пензе [83]. В 2009 г. авторы этого доклада зарегистрировали свою идею в Роспатенте в виде «Лабораторного практикума по нейросетевым технологиям» [84], в котором студентам предлагалось, используя свои скромные медицинские знания, сформировать датасет и с помощью него обучить нейронную сеть ставить диагнозы заболеваний: «пневмония» и «острое респираторное заболевание». Лабораторный практикум оказался таким удачным, что уже более 15 лет используется в учебном процессе некоторыми российскими вузами. Тем не менее российские исследователи не используют понятие «синтетические данные». В итоге, когда возникает проблема поиска публикаций по этой теме, поисковые системы выдают крайне скудные результаты. Однако, стоит только ввести в строку поискового запроса термин synthetic data, так сразу же открывается океан информации. То есть исследователи должны знать этот устоявшийся общеупотребимый термин, чтобы не страдать от отсутствия информации. Именно желание помочь в поиске информации в области синтетических данных и было основным мотивирующим фактором для написания настоящей работы.

Анализируя первоисточники, видим, что большая часть публикаций сделана в последние три года, при этом каждый год количество публикаций увеличивается в разы по сравнению с предыдущим годом. За первую половину 2022 г. уже опубликовано столько же работ, сколько за весь предыдущий 2021 г. Рост количества публикаций продолжается так же, как продолжается и расширение сфер применения синтетических данных, и, как уже упоминалось в начале статьи, по оценкам Gartner к 2030 г. количество синтетических данных для обучения нейросетевых моделей превысит количество реальных данных. Синтетические данные дешевы в производстве, они доступны в неограниченном количестве, они решают проблему разреженности или несбалансированности данных, их можно использовать, когда реальные данные невозможно получить, и, наконец, они обеспечивают полную конфиденциальность, что облегчает обмен данными между исследователями, что еще более способствует развитию технологий искусственного интеллекта. Таким образом, синтетические данные становятся необходимым элементом в современных нейросетевых технологиях. Для успешного развития систем искусственного интеллекта сегодня уже недостаточно просто научиться использовать синтетические данные, необходимо выполнять научные исследования в области разработки новых алгоритмов и методов их генерации, а также проводить технологические исследования, на базе которых создавать современные инструменты и программные комплексы, которые позволят обеспечить исследователей достаточным количеством обучающих данных для постоянного совершенствования нейросетевых технологий.

Список литературы

1. Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation / R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham, M.A. Zissman // In: Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00. IEEE Comput. Soc. – 2000. – P. 12–26. DOI: 10.1109/DISCEX.2000.821506

2. Lundin E., Kvarnström H., Jonsson E.A. Synthetic Fraud Data Generation Methodology // Deng Robertand Bao, Fengand Zhou Jianyingand, and Qing Sihan (eds.) Information and Communications Security. Springer Berlin Heidelberg, Berlin, Heidelberg. – 2002. – P. 265–277. DOI: 10.1007/3-540-36159-6_23

3. Learning Deep Models from Synthetic Data for Extracting Dolphin Whistle Contours / P. Li, X. Liua, K.J. Palmer, E. Fleishman, D. Gillespie, E.-M. Nosal, Y. Shiu, H. Klinck, D. Cholewiak, T. Helble, M.A. Roch. – 2020. DOI: 10.48550/ARXIV.2005.08894

4. Lombardo, J. Method for Generation and Distribution of Synthetic Medical Record Data for Evaluation of Disease-Monitoring Systems / J. Lombardo, L.A. Moniz // Johns Hopkins APL Technical Digest (Applied Physics Laboratory). – 2008. – Vol. 27.

5. Construction and Validation of Synthetic Electronic Medical Records / L. Moniz, A.L. Buczak, L. Hung, S. Babin, M. Dorko, J. Lombardo // Online J Public Health Inform. – 2009. – Vol. 1. DOI: 10.5210/ojphi.v1i1.2720

6. Buczak A.L., Babin S., Moniz L. Data-driven approach for creating synthetic electronic medical records // BMC Med Inform Decis Mak. – 2010. – Vol. 10, iss. 1. DOI: 10.1186/1472-6947-10-59

7. Jin C., Rinard M.C. Learning From Context-Agnostic Synthetic Data // CoRR. – 2020. – abs/2005.14707

8. McKenna R., Miklau G., Sheldon D. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data // CoRR. – 2021. – abs/2108.04978

9. Noise-Aware Statistical Inference with Differentially Private Synthetic Data / O. Räisä, J. Jälkö, S. Kaski, A. Honkela // arXiv. – 2022. – abs/2205.14485. DOI: 10.48550/ARXIV.2205.14485

10. Awan J., Cai Z. One Step to Efficient Synthetic Data // arXiv. – 2020. – bs/2006.02397. DOI: 10.48550/ARXIV.2006.02397

11. Goetz J., Tewari A. Federated Learning via Synthetic Data // CoRR. – 2020. – abs/2008.04489

12. FedSynth: Gradient Compression via Synthetic Data in Federated Learning / S. Hu, J. Goetz, K. Malik, H. Zhan, Z. Liu, Y. Liu // arXiv. – 2022. – abs/2204.01273. DOI: 10.48550/ARXIV.2204.01273

13. SMOTE: Synthetic Minority Over-sampling Technique / N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer // Journal of Artificial Intelligence Research. – 2002. – Vol. 16. – P. 321–357. DOI: 10.1613/jair.953

14. Mukherjee M., Khushi M. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features // Applied System Innovation. – 2021. – Vol. 4, iss. 1. – Art. 18. DOI: 10.3390/asi4010018.

15. A Method for Handling Multi-class Imbalanced Data by Geometry based Information Sampling and Class Prioritized Synthetic Data Generation (GICaPS) / A. Majumder, S. Dutta, S. Kumar, L. Behera // CoRR. – 2020. – abs/2010.05155

16. Gonsior J., Thiele M., Lehner W. ImitAL: Learning Active Learning Strategies from Synthetic Data //. CoRR. – 2021. – abs/2108.07670

17. Dataset Condensation via Efficient Synthetic-Data Parameterization / J.-H. Kim, J. Kim, S.J. Oh, S. Yun, H. Song, J. Jeong, J.-W. Ha, H.O. Song // arXiv. – 2022. – abs/2205.14959. DOI: 10.48550/ARXIV.2205.14959

18. Effective Use of Synthetic Data for Urban Scene Semantic Segmentation / F.S. Saleh, M.S. Aliakbarian, M. Salzmann, L. Petersson, J.M. Alvarez // CoRR. – 2018. – abs/1807.06132

19. Mason K., Vejdan S., Grijalva S. An "On The Fly" Framework for Efficiently Generating Synthetic Big Data Sets // CoRR. – 2019. – abs/1903.06798

20. Condrea F., Ivan V.-A., Leordeanu M. In Search of Life: Learning from Synthetic Data to Detect Vital Signs in Videos // CoRR. – 2020. – abs/2004.07691

21. Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data / H.P. Das, R. Tran, J. Singh, X. Yue, G. Tison, A.L. Sangiovanni-Vincentelli, C.J. Spanos // CoRR. – 2021. – abs/2109.06486

22. Рабчевский А.Н., Ашихмин Е.Г., Рабчевский Е.А. Моделирование структуры пропаганды протестного движения в социальных сетях с помощью графового анализа и нейросетевых технологий. – текст: непосредственный // Математическое и компьютерное моделирование: сборник материалов IX Международной научной конференции, посвященной 85-летию профессора В.И. Потапова. – Омск, 2021. – С. 273–276.

23. Rabchevsky A., Yasnitsky L., Zayakin V. Comparison of methods for identifying user roles in online social networks // Applied Mathematics and Control Sciences. – 2021. Vol. 2. – P. 93–111. DOI: 10.15593/2499-9873/2021.2.06

24. Rabchevskiy A.N., Yasnitskiy L.N. Creating and Using Synthetic Data for Neural Network Training, Using the Creation of a Neural Network Classifier of Online Social Network User Roles as an Example // Digital Science. DSIC 2021. Lecture Notes in Networks and Systems, Springer, Cham. – 2022. – Vol. 381. – P. 412–421. DOI: 10.1007/978-3-030-93677-8_36

25. Generation of synthetic training data for object detection in piles / E. Buls, R. Kadikis, R. Cacurs, J. Ārents // D.P. Nikolaev, P. Radeva, A. Verikas, J. Zhou (eds.) Eleventh International Conference on Machine Vision (ICMV 2018), SPIE. – 2019. – P. 105. DOI: 10.1117/12.2523203

26. An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance Detection / S. Hinterstoisser, O. Pauly, H. Heibel, M. Marek, M. Bokeloh // CoRR. – 2019. – abs/1902.09967

27. Dina A.S., Siddique A.B., Manivannan D. Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks // arXiv. – 2022. – abs/2204.00144. DOI: 10.48550/ARXIV.2204.00144

28. Charitou C., Dragicevic S., d'Avila Garcez A. Synthetic Data Generation for Fraud Detection using GANs // CoRR. – 2021. – abs/2109.12546

29. Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study / C. Little, M. Elliot, R. Allmendinger, S.S. Samani // arXiv. – 2021. – abs/2112.01925. DOI: 10.48550/ARXIV.2112.01925

30. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification / M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan // Neurocomputing. – 2018. – Vol. 321. – P. 321–331. DOI: 10.1016/j.neucom.2018.09.013

31. GAN-based synthetic brain MR image generation / C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, H. Nakayama // 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). – 2018. – P. 734–738.

32. Constrained Generative Adversarial Network Ensembles for Sharable Synthetic Data Generation / E. Dikici, L.M. Prevedello, M. Bigelow, R.D. White, B.S. Erdal // arXiv. – 2020. – abs/2003.00086. DOI: 10.48550/ARXIV.2003.00086

33. Using Synthetic Data to Enhance the Accuracy of Fingerprint-Based Localization: A Deep Learning Approach / M. Nabati, H. Navidan, R. Shahbazian, S.A. Ghorashi, D. Windridge // IEEE Sens Lett. – 2020. – Vol. 4. – P. 1–4. DOI: 10.1109/lsens.2020.2971555

34. Synthetic Data Generation and Adaption for Object Detection in Smart Vending Machines / K. Wang, F. Shi, W. Wang, Y. Nan, S. Lian // CoRR. – 2019. – abs/1904.12294

35. Synthetic Data and Hierarchical Object Detection in Overhead Imagery / N. Clement, A. Schoen, A.P. Boedihardjo, A. Jenkins // CoRR. – 2021. – abs/2102.00103

36. Jordon J., Yoon J., van der Schaar M. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees // In: ICLR. – 2019.

37. Arnold C. Releasing differentially private synthetic micro-data with bayesian gans. – 2018.

38. Li M., Zhuang D., Chang J.M. MC-GEN:Multi-level Clustering for Private Synthetic Data Generation // arXiv. – 2022. – abs/2205.14298. DOI: 10.48550/ARXIV.2205.14298

39. Incentivizing Collaboration in Machine Learning via Synthetic Data Rewards / S.S. Tay, X. Xu, C.S. Foo, B.K.H. Low // CoRR. – 2021. – abs/2112.09327

40. Liu T., Vietri G., Wu Z.S. Iterative Methods for Private Synthetic Data: Unifying Framework and New Methods // CoRR. – 2021. – abs/2106.07153

41. Low Light Video Enhancement using Synthetic Data Produced with an Intermediate Domain Mapping / D. Triantafyllidou, S. Moran, S. McDonagh, S. Parisot, G. Slabaugh // arXiv. – 2020. –abs/2007.09187. – DOI: 10.48550/ARXIV.2007.09187

42. Adapting deep generative approaches for getting synthetic data with realistic marginal distributions / K. Farhadyar, F. Bonofiglio, D. Zoeller, H. Binder // atXiv. – 2021. – abs/2105.06907. DOI: 10.48550/ARXIV.2105.06907

43. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks / B. van Breugel, T. Kyono, J. Berrevoets, M. van der Schaar // CoRR. – 2021. – abs/2110.12884

44. Graham P., Penny R. Multiply Imputed Synthetic Data Files // Official Statistics Research Series, Statistics New Zealand. – 2007. – Vol. 1.

45. Boedihardjo M., Strohmer T., Vershynin R. Private sampling: a noiseless approach for generating differentially private synthetic data // CoRR. – 2021. – abs/2109.14839

46. Boedihardjo M., Strohmer T., Vershynin R. Covariance's Loss is Privacy's Gain: Computationally Efficient, Private and Accurate Synthetic Data // CoRR. – 2021. – abs/2107.05824

47. Kamthe S., Assefa S., Deisenroth M. Copula Flows for Synthetic Data Generation // arXiv. – 2021. – abs/2101.00598. DOI: 10.48550/ARXIV.2101.00598

48. Li Z., Zhao Y., Fu J. SYNC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources // arXiv. – 2020. – abs/2009.09471. DOI: 10.48550/ARXIV. 2009.09471

49. End-to-End Synthetic Data Generation for Domain Adaptation of Question Answering Systems / S. Shakeri, C.N. dos Santos, H. Zhu, P. Ng, F. Nan, Z. Wang, R. Nallapati, B. Xiang // CoRR. – 2020. – abs/2010.06028

50. Bousquet O., Livni R., Moran S. Synthetic Data Generators: Sequential and Private // arXiv. – 2019. – abs/1902.03468. DOI: 10.48550/ARXIV.1902.03468

51. Exploring Invariances in Deep Convolutional Neural Networks Using Synthetic Images / X. Peng, B. Sun, K. Ali, K. Saenko // CoRR. – 2014. – abs/1412.7122

52. Transfer Learning from Synthetic to Real Images Using Variational Autoencoders for Precise Position Detection / T. Inoue, S. Choudhury, G. de Magistris, S. Dasgupta // 2018

25th IEEE International Conference on Image Processing (ICIP). – 2018. – P. 2725–2729. DOI: 10.1109/ICIP.2018.8451064

53. Learning to Augment Synthetic Images for Sim2Real Policy Transfer / A. Pashevich, R. Strudel, I. Kalevatykh, I. Laptev, C. Schmid // CoRR. – 2019. – abs/1903.07740. DOI: 10.48550/arXiv.1903.07740

54. AutoSimulate: (Quickly) Learning Synthetic Data Generation / H.S. Behl, A.G. Baydin, R. Gal, P.H.S. Torr, V. Vineet // CoRR. – 2020. – abs/2008.08424.

55. ProcSy: Procedural Synthetic Dataset Generation Towards Influence Factor Studies Of Semantic Segmentation Networks / S. Khan, B. Phan, R. Salay, K. Czarnecki // CVPR Workshops. – 2019.

56. Illumination Invariant Camera Localization Using Synthetic Images / S. Shoman, T. Mashita, A. Plopski, P. Ratsamee, Y. Uranishi, H. Takemura // 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). – IEEE. – 2018. – P. 143–144. DOI: 10.1109/ISMAR-Adjunct.2018.00053

57. Rozantsev A., Lepetit V., Fua P. On rendering synthetic images for training an object detector // Computer Vision and Image Understanding. – 2015. – Vol. 137. – P. 24–37. DOI: 10.1016/j.cviu.2014.12.006

58. Synthetic Data Are as Good as the Real for Association Knowledge Learning in Multi-object Tracking / Y. Liu, Z. Wang, X. Zhou, L. Zheng // CoRR. – 2021. – abs/2106.16100

59. Automatic Generation of Synthetic LiDAR Point Clouds for 3-D Data Analysis / F. Wang, Y. Zhuang, H. Gu, H. Hu // IEEE Trans Instrum Meas. – 2019. – Vol. 68. – P. 2671–2673. DOI: 10.1109/TIM.2019.2906416

60. Learning how to analyse crowd behaviour using synthetic data / A.R. Khadka, M.M. Oghaz, W. Matta, M. Cosentino, P. Remagnino, V. Argyriou // Proceedings of the 32nd International Conference on Computer Animation and Social Agents. – ACM. – New York. – NY. – USA. – 2019. – P. 11–14. DOI: 10.1145/3328756.3328773

61. Synthetic Data Generation for Deep Learning of Underwater Disparity Estimation / E.A. Olson, C. Barbalata, J. Zhang, K.A. Skinner, M. Johnson-Roberson // OCEANS 2018 MTS/IEEE Charleston. – 2018. – P. 1–6.

62. Sun S., Shi H., Wu Y. A survey of multi-source domain adaptation // Information Fusion. – 2015. – Vol. 24. – P. 84–92. DOI: 10.1016/j.inffus.2014.12.003

63. Ren Z., Lee Y.J. Cross-Domain Self-Supervised Multi-task Feature Learning Using Synthetic Imagery // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2018. – P. 762–771. DOI: 10.1109/CVPR.2018.00086

64. Effect of Kinematics and Fluency in Adversarial Synthetic Data Generation for ASL Recognition with RF Sensors / M.M. Rahman, E. Malaia, A.C. Gurbuz, D.J. Griffin, C. Crawford, S. Gurbuz // IEEE Trans Aerosp Electron Syst. – 2022. – Vol. 1. DOI: 10.1109/taes.2021.3139848

65. Alkhalifah T., Wang H., Ovcharenko O. MLReal: Bridging the gap between training on synthetic data and real data applications in machine learning // arXiv. – 2021. – abs/2109.05294. DOI: 10.48550/ARXIV.2109.05294

66. Learning from Synthetic Data for Opinion-free Blind Image Quality Assessment in the Wild / Z. Wang, Z.-R. Tang, Z. Yu, J. Zhang, Y. Fang // CoRR. – 2021. – abs/2106.14076

67. Meta-Sim: Learning to Generate Synthetic Datasets / A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, A. Fidler // CoRR. – 2019. – abs/1904.11621

68. Stein G.J., Roy N. GeneSIS-RT: Generating Synthetic Images for training Secondary Real-world Tasks // CoRR. – 2017. – abs/1710.04280. DOI: 10.48550/arXiv.1710.04280

69. S^3Net: Semantic-Aware Self-supervised Depth Estimation with Monocular Videos and Synthetic Data / B. Cheng, I.S. Saggu, R. Shah, G. Bansal, D. Bharadia // CoRR. – 2020. – abs/2007.14511

70. PeopleSansPeople: A Synthetic Data Generator for Human-Centric Computer Vision / S.E. Ebadi, Y.-C. Jhang, A. Zook, S. Dhakad, A. Crespi, P. Parisi, S. Borkman, J. Hogins, S. Ganguly // CoRR. – 2021. – abs/2112.09290

71. Hart K.M., Goodman A.B., O'Shea R.P. Automatic Generation of Machine Learning Synthetic Data Using ROS // CoRR. – 2021. – abs/2106.04547

72. UnrealROX+: An Improved Tool for Acquiring Synthetic Data from Virtual 3D Environments / P. Martinez-Gonzalez, S. Oprea, J.A. Castro-Vargas, A. Garcia-Garcia, S. Orts-Escolano, J.G. Rodrguez, M. Vincze // CoRR. – 2021. – abs/2104.11776

73. Jin C., Rinard M.C. Learning From Context-Agnostic Synthetic Data // CoRR. – 2020. – abs/2005.14707

74. Baek K., Shim H. Commonality in Natural Images Rescues GANs: Pretraining GANs with Generic and Privacy-free Synthetic Data // arXiv. – 2022. – abs/2204.04950. DOI: 10.48550/ARXIV.2204.04950

75. Deep Learning based Food Instance Segmentation using Synthetic Data / D. Park, J. Lee, J. Lee, K. Lee // CoRR. – 2021. – abs/2107.07191

76. Raab G.M., Nowok B., Dibben C. Assessing, visualizing and improving the utility of synthetic data // arXiv. – 2021. – abs/2109.12717. DOI: 10.48550/ARXIV.2109.12717

77. Unity Perception: Generate Synthetic Data for Computer Vision / S. Borkman, A. Crespi, S. Dhakad, S. Ganguly, J. Hogins, Y.-C. Jhang, M. Kamalzadeh, B. Li, S. Leal, P. Parisi, C. Romero, W. Smith, A. Thaman, S. Warren, N. Yadav // CoRR. – 2021. – abs/2107.04259

78. ElderSim: A Synthetic Data Generation Platform for Human Action Recognition in Eldercare Applications / H. Hwang, C. Jang, G. Park, J. Cho, I.-J. Kim // arXiv. – 2020. – abs/2010.14742. – DOI: 10.48550/ARXIV.2010.14742

79. Dilmegani G. Top 20 Synthetic Data Use Cases & Applications in 2023: сайт [Электронный ресурс]. – URL: https://research.aimultiple.com/synthetic-data-use-cases (дата обращения: 13.09.2023).

80. Dilmegani G. The Ultimate Guide to Synthetic Data: Uses, Benefits & Tools: сайт [Электронный ресурс]. – URL: https://research.aimultiple.com/synthetic-data-tools (дата обращения: 13.09.2023).

81. Dilmegani G. The Ultimate Guide to Synthetic Data in 2023: сайт [Электронный ресурс]. – URL: https://research.aimultiple.com/synthetic-data/ (дата обращения: 13.09.2023).

82. Dilmegani G. Synthetic Data Generation: Techniques, Best Practices & Tools: сайт [Электронный ресурс]. – URL: https://research.aimultiple.com/synthetic-data-generation/ (дата обращения: 13.09.2023).

83. Черепанов Ф.М., Ясницкий Л.Н. Лабораторный практикум по нейросетевым технологиям. – текст: непосредственный. // Перспективные технологии искусственного интеллекта: сборник трудов международной научно-практической конференции (Пенза, Пензенский ун-т, Научный Совет РАН по методологии искусственного интеллекта, 1-6 июля 2008 г.) / Пенз. ун-т. – Пенза. – 2008. – С. 128–130.

84. Черепанов Ф.М., Ясницкий Л.Н. Лабораторный практикум по нейросетевым технологиям: свидетельство о государственной регистрации программы для ЭВМ № 2009611544. Заявка № 2009610226. Зарегистрировано в Реестре программ для ЭВМ 12 марта 2009 г.

References

1. Lippmann R.P., Fried D.J., Graf I., Haines J.W., Kendall K.R., McClung D., Weber D., Webster S.E., Wyschogrod D., Cunningham R.K., Zissman M.A. Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation. *In: Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00. IEEE Comput. Soc.,* 2000, pp. 12–26. DOI: 10.1109/DISCEX.2000.821506

2. Lundin E., Kvarnström H., Jonsson E. A Synthetic Fraud Data Generation Methodology. *In: Deng Robertand Bao, Fengand Zhou Jianyingand, and Qing Sihan (eds.) Information and Communications Security*. *Springer Berlin Heidelberg, Berlin, Heidelberg,* 2002, pp. 265–277. DOI: 10.1007/3-540-36159-6_23

3. Li P., Liua X., Palmer K.J., Fleishman E., Gillespie D., Nosal E.-M., Shiu Y., Klinck H., Cholewiak D., Helble T., Roch M.A. Learning Deep Models from Synthetic Data for Extracting Dolphin Whistle Contours. 2020. DOI: 10.48550/ARXIV.2005.08894

4. Lombardo J., Moniz L. A Method for Generation and Distribution of Synthetic Medical Record Data for Evaluation of Disease-Monitoring Systems. *Johns Hopkins APL Technical Digest (Applied Physics Laboratory)*, 2008, vol. 27.

5. Moniz L., Buczak A.L., Hung L., Babin S., Dorko M., Lombardo J. Construction and Validation of Synthetic Electronic Medical Records. *Online J Public Health Inform*, 2009, vol.1. DOI: 10.5210/ojphi.v1i1.2720

6. Buczak A.L., Babin S., Moniz L. Data-driven approach for creating synthetic electronic medical records. *BMC Med Inform Decis Mak,* 2010, no.10, vol.59. DOI: 10.1186/1472-6947-10-59

7. Jin C., Rinard M.C. Learning From Context-Agnostic Synthetic Data. *CoRR*, 2020. abs/2005.14707

8. McKenna R., Miklau G., Sheldon D. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *CoRR*, 2021. abs/2108.04978

9. Räisä O., Jälkö J., Kaski S., Honkela A. Noise-Aware Statistical Inference with Differentially Private Synthetic Data. *arXiv*, 2022, abs/2205.14485. DOI: 10.48550/ARXIV.2205.14485

10. Awan J., Cai Z. One Step to Efficient Synthetic Data. *arXiv*, 2020, abs/2006.02397. DOI: 10.48550/ARXIV.2006.02397

11. Goetz J., Tewari A. Federated Learning via Synthetic Data. *CoRR,* 2020. abs/2008.04489

12. Hu S., Goetz J., Malik K., Zhan H., Liu Z., Liu Y. FedSynth: Gradient Compression via Synthetic Data in Federated Learning. *arXiv*, 2022, abs/2204.01273. DOI 10.48550/ARXIV.2204.01273

13. Chawla N.V., Bowyer K.W., Hall, L.O., Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002, vol. 16, pp. 321–357. DOI: 10.1613/jair.953.

14. Mukherjee M., Khushi M. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, 2021, vol. 4, iss. 1, art. 18. DOI 10.3390/asi4010018.

15. Majumder A., Dutta S., Kumar S., Behera L. A Method for Handling Multi-class Imbalanced Data by Geometry based Information Sampling and Class Prioritized Synthetic Data Generation (GICaPS). *CoRR*, 2020. abs/2010.05155

16. Gonsior J., Thiele M., Lehner W. ImitAL: Learning Active Learning Strategies from Synthetic Data. *CoRR*, 2021. abs/2108.07670

17. Kim J.-H., Kim J., Oh S.J., Yun S., Song H., Jeong J., Ha J.-W., Song H.O. Dataset Condensation via Efficient Synthetic-Data Parameterization. *arXiv*, 2022, abs/2205.14959. DOI: 10.48550/ARXIV.2205.14959

18. Saleh F.S., Aliakbarian M.S., Salzmann M., Petersson L., Alvarez J.M. Effective Use of Synthetic Data for Urban Scene Semantic Segmentation. *CoRR*, 2018. abs/1807.06132

19. Mason K., Vejdan S., Grijalva S. An “On The Fly” Framework for Efficiently Generating Synthetic Big Data Sets. *CoRR*, 2019. abs/1903.06798

20. Condrea F., Ivan V.-A., Leordeanu M. In Search of Life: Learning from Synthetic Data to Detect Vital Signs in Videos. *CoRR*, 2020. abs/2004.07691

21. Das H.P., Tran R., Singh J., Yue X., Tison G., Sangiovanni-Vincentelli A.L., Spanos C.J. Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data. *CoRR*, 2021. abs/2109.06486

22. Rabchevskij A.N., Ashihmin E.G., Rabchevskij E.A. Modelirovanie struktury propagandy protestnogo dvizheniya v social'nyh setyah s pomoshch'yu grafovogo analiza i nejrosetevyh tekhnologij. *Matematicheskoe i komp'yuternoe modelirovanie. Sbornik materialov IX Mezhdunarodnoj nauchnoj konferencii, posvyashchennoj 85-letiyu professora V.I. Potapova, Omsk,* 2021, pp. 273–276.

23. Rabchevsky A., Yasnitsky L., Zayakin V. Comparison of methods for identifying user roles in online social networks. *Applied Mathematics and Control Sciences*, 2021, pp. 93–111. DOI: 10.15593/2499-9873/2021.2.06

24. Rabchevskiy A.N., Yasnitskiy L.N. Creating and Using Synthetic Data for Neural Network Training, Using the Creation of a Neural Network Classifier of Online Social Network User Roles as an Example. *Digital Science. DSIC 2021. Lecture Notes in Networks and Systems, Springer, Cham.,* 2022, vol. 381, pp. 412–421. DOI: 10.1007/978-3-030-93677-8_36

25. Buls E., Kadikis R., Cacurs R., Ārents J. Generation of synthetic training data for object detection in piles. *In: Nikolaev, D.P., Radeva, P., Verikas, A., and Zhou, J. (eds.) Eleventh International Conference on Machine Vision (ICMV 2018), SPIE,* 2019, p.105. DOI: 10.1117/ 12.2523203

26. Hinterstoisser S., Pauly O., Heibel H., Marek M., Bokeloh M. An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance Detection. *CoRR*, 2019. abs/1902.09967

27. Dina A.S., Siddique A.B., Manivannan D. Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks. *arXiv*, 2022, abs/2204.00144. DOI: 10.48550/ARXIV.2204.00144

28. Charitou C., Dragicevic S., d’Avila Garcez A. Synthetic Data Generation for Fraud Detection using GANs. *CoRR, 2021.* abs/2109.12546

29. Little C., Elliot M., Allmendinger R., Samani S.S. Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. *arXiv*, 2021, abs/2112.01925. DOI: 10.48550/ARXIV.2112.01925

30. Frid-Adar M., Diamant I., Klang E., Amitai M., Goldberger J., Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 2018, vol.321, pp.321–331. DOI: 10.1016/j.neucom.2018.09.013

31. Han C., Hayashi H., Rundo L., Araki R., Shimoda W., Muramatsu S., Furukawa Y., Mauri G., Nakayama H. GAN-based synthetic brain MR image generation. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018),* 2018, pp. 734–738.

32. Dikici E., Prevedello L.M., Bigelow M., White R.D., Erdal B.S. Constrained Generative Adversarial Network Ensembles for Sharable Synthetic Data Generation. *arXiv*, 2020, abs/2003.00086. DOI: 10.48550/ARXIV.2003.00086

33. Nabati M., Navidan H., Shahbazian R., Ghorashi S.A., Windridge D. Using Synthetic Data to Enhance the Accuracy of Fingerprint-Based Localization: A Deep Learning Approach. *IEEE Sens Lett*., 2020, vol.4, pp.1–4. DOI: 10.1109/lsens.2020.2971555

34. Wang K., Shi F., Wang W., Nan Y., Lian S. Synthetic Data Generation and Adaption for Object Detection in Smart Vending Machines. *CoRR*, 2019, abs/1904.12294

35. Clement N., Schoen A., Boedihardjo A.P., Jenkins A. Synthetic Data and Hierarchical Object Detection in Overhead Imagery. *CoRR*, 2021. abs/2102.00103

36. Jordon J., Yoon J., van der Schaar M. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. *In: ICLR*, 2019.

37. Arnold C. Releasing differentially private synthetic micro-data with bayesian gans. 2018.

38. Li M., Zhuang D., Chang J.M. MC-GEN:Multi-level Clustering for Private Synthetic Data Generation. *arXiv*, 2022, abs/2205.14298. DOI: 10.48550/ARXIV.2205.14298

39. Tay S.S., Xu X., Foo C.S., Low B.K.H. Incentivizing Collaboration in Machine Learning via Synthetic Data Rewards. *CoRR*, 2021. abs/2112.09327

40. Liu T., Vietri G., Wu Z.S. Iterative Methods for Private Synthetic Data: Unifying Framework and New Methods. *CoRR,* 2021. abs/2106.07153

41. Triantafyllidou D., Moran S., McDonagh S., Parisot S., Slabaugh G. Low Light Video Enhancement using Synthetic Data Produced with an Intermediate Domain Mapping. *arXiv*, 2020, abs/2007.09187. DOI: 10.48550/ARXIV.2007.09187

42. Farhadyar K., Bonofiglio F., Zoeller D., Binder H. Adapting deep generative approaches for getting synthetic data with realistic marginal distributions. *arXiv*, 2021, abs/2105.06907. DOI: 10.48550/ARXIV.2105.06907

43. van Breugel B., Kyono T., Berrevoets J., van der Schaar M. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. *CoRR,* 2021. abs/2110.12884

44. Graham P., Penny R. Multiply Imputed Synthetic Data Files. *Official Statistics Research Series, Statistics New Zealand,* 2007, vol.1.

45. Boedihardjo M., Strohmer T., Vershynin R. Private sampling: a noiseless approach for generating differentially private synthetic data. *CoRR*, 2021. abs/2109.14839

46. Boedihardjo M., Strohmer T., Vershynin R. Covariance's Loss is Privacy's Gain: Computationally Efficient, Private and Accurate Synthetic Data. *CoRR*, 2021. abs/2107.05824

47. Kamthe S., Assefa S., Deisenroth M. Copula Flows for Synthetic Data Generation. *arXiv*, 2021, abs/2101.00598. DOI: 10.48550/ARXIV.2101.00598

48. Li Z., Zhao Y., Fu J. SYNC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources. *arXiv*, 2020, abs/2009.09471. DOI: 10.48550/ARXIV.2009.09471

49. Shakeri S., dos Santos C.N., Zhu H., Ng P., Nan F., Wang Z., Nallapati R., Xiang B. End-to-End Synthetic Data Generation for Domain Adaptation of Question Answering Systems. *CoRR*, 2020. abs/2010.06028

50. Bousquet O., Livni R., Moran S. Synthetic Data Generators: Sequential and Private. *arXiv*, 2019, abs/1902.03468. DOI: 10.48550/ARXIV.1902.03468

51. Peng X., Sun B., Ali K., Saenko K. Exploring Invariances in Deep Convolutional Neural Networks Using Synthetic Images. *CoRR,* 2014. abs/1412.7122

52. Inoue T., Choudhury S., de Magistris G., Dasgupta S. Transfer Learning from Synthetic to Real Images Using Variational Autoencoders for Precise Position Detection. *In: 2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2725–2729. DOI: 10.1109/ICIP.2018.8451064

53. Pashevich A., Strudel R., Kalevatykh I., Laptev I., Schmid C. Learning to Augment Synthetic Images for Sim2Real Policy Transfer. *CoRR,* 2019, abs/1903.07740. DOI: 10.48550/arXiv.1903.07740

54. Behl H.S., Baydin A.G., Gal R., Torr P.H.S., Vineet V. AutoSimulate: (Quickly) Learning Synthetic Data Generation. *CoRR,* 2020. abs/2008.08424

55. Khan S., Phan B., Salay R., Czarnecki K. ProcSy: Procedural Synthetic Dataset Generation Towards Influence Factor Studies Of Semantic Segmentation Networks. *CVPR Workshops,* 2019.

56. Shoman S., Mashita T., Plopski A., Ratsamee P., Uranishi Y., Takemura H. Illumination Invariant Camera Localization Using Synthetic Images. *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE*, 2018, pp.143–144. DOI: 10.1109/ISMAR-Adjunct.2018.00053

57. Rozantsev A., Lepetit V., Fua P. On rendering synthetic images for training an object detector. *Computer Vision and Image Understanding*, 2015, vol.137, pp. 24–37. DOI: 10.1016/j.cviu.2014.12.006

58. Liu Y., Wang Z., Zhou X., Zheng L. Synthetic Data Are as Good as the Real for Association Knowledge Learning in Multi-object Tracking. *CoRR*, 2021. abs/2106.16100

59. Wang F., Zhuang Y., Gu H., Hu H. Automatic Generation of Synthetic LiDAR Point Clouds for 3-D Data Analysis. *IEEE Trans Instrum Meas*, 2019, vol.68, pp.2671–2673. DOI: 10.1109/TIM.2019.2906416

60. Khadka A.R., Oghaz M.M., Matta W., Cosentino M., Remagnino P., Argyriou V. Learning how to analyse crowd behaviour using synthetic data. *Proceedings of the 32nd International Conference on Computer Animation and Social Agents, ACM, New York, NY, USA,* 2019, pp.11–14. DOI: 10.1145/3328756.3328773

61. Olson E.A., Barbalata C., Zhang J., Skinner K.A., Johnson-Roberson M. Synthetic Data Generation for Deep Learning of Underwater Disparity Estimation. *OCEANS 2018 MTS/IEEE Charleston*, 2018, pp.1–6.

62. Sun S., Shi H., Wu Y. A survey of multi-source domain adaptation. Information Fusion, 2015, vol.24, pp.84–92. DOI: 10.1016/j.inffus.2014.12.003

63. Ren Z., Lee Y.J. Cross-Domain Self-Supervised Multi-task Feature Learning Using Synthetic Imagery. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp.762–771. DOI: 10.1109/CVPR.2018.00086

64. Rahman M.M., Malaia E., Gurbuz A.C., Griffin D.J., Crawford C., Gurbuz S. Effect of Kinematics and Fluency in Adversarial Synthetic Data Generation for ASL Recognition with RF Sensors. *IEEE Trans Aerosp Electron Syst*, 2022, vol.1. DOI: 10.1109/taes.2021.3139848

65. Alkhalifah T., Wang H., Ovcharenko O. MLReal: Bridging the gap between training on synthetic data and real data applications in machine learning. *arXiv*, 2021, abs/2109.05294. DOI: 10.48550/ARXIV.2109.05294

66. Wang Z., Tang Z.-R., Yu Z., Zhang J., Fang Y. Learning from Synthetic Data for Opinion-free Blind Image Quality Assessment in the Wild. *CoRR*, 2021. abs/2106.14076

67. Kar A., Prakash A., Liu M.-Y., Cameracci E., Yuan J., Rusiniak M., Acuna D., Torralba A., Fidler S. Meta-Sim: Learning to Generate Synthetic Datasets. *CoRR,* 2019. abs/1904.11621

68. Stein G.J., Roy N. GeneSIS-RT: Generating Synthetic Images for training Secondary Real-world Tasks. *CoRR*, 2017, abs/1710.04280. DOI: 10.48550/arXiv.1710.04280

69. Cheng B., Saggu I.S., Shah R., Bansal G., Bharadia D. S^3Net: Semantic-Aware Self-supervised Depth Estimation with Monocular Videos and Synthetic Data. *CoRR*, 2020. abs/2007.14511

70. Ebadi S.E., Jhang Y.-C., Zook A., Dhakad S., Crespi A., Parisi P., Borkman S., Hogins J., Ganguly S. PeopleSansPeople: A Synthetic Data Generator for Human-Centric Computer Vision. *CoRR*, 2021. abs/2112.09290

71. Hart K.M., Goodman A.B., O'Shea R.P. Automatic Generation of Machine Learning Synthetic Data Using ROS. *CoRR*, 2021. abs/2106.04547

72. Martinez-Gonzalez P., Oprea S., Castro-Vargas J.A., Garcia-Garcia A., Orts-Escolano S., Rodrguez J.G., Vincze M. UnrealROX+: An Improved Tool for Acquiring Synthetic Data from Virtual 3D Environments. *CoRR*, 2021. abs/2104.11776

73. Jin C., Rinard M.C. Learning From Context-Agnostic Synthetic Data. *CoRR*, 2020. abs/2005.14707

74. Baek K., Shim H. Commonality in Natural Images Rescues GANs: Pretraining GANs with Generic and Privacy-free Synthetic Data. *arXiv*, 2022, abs/2204.04950, DOI: 10.48550/ARXIV.2204.04950

75. Park D., Lee J., Lee J., Lee K. Deep Learning based Food Instance Segmentation using Synthetic Data. *CoRR*, 2021. abs/2107.07191

76. Raab G.M., Nowok B., Dibben C. Assessing, visualizing and improving the utility of synthetic data. *arXiv*, 2021, abs/2109.12717. DOI: 10.48550/ARXIV.2109.12717

77. Borkman S., Crespi A., Dhakad S., Ganguly S., Hogins J., Jhang Y.-C., Kamalzadeh M., Li B., Leal S., Parisi P., Romero C., Smith W., Thaman A., Warren S., Yadav N. Unity Perception: Generate Synthetic Data for Computer Vision. *CoRR*, 2021. abs/2107.04259

78. Hwang H., Jang C., Park G., Cho J., Kim I.-J. ElderSim: A Synthetic Data Generation Platform for Human Action Recognition in Eldercare Applications. *arXiv*, 2020, abs/2010.14742. DOI: 10.48550/ARXIV.2010.14742

79. Dilmegani G. Top 20 Synthetic Data Use Cases & Applications in 2023, available at: https://research.aimultiple.com/synthetic-data-use-cases/ (accessed 13 September 2023)

80. Dilmegani G. The Ultimate Guide to Synthetic Data: Uses, Benefits & Tools, available at: https://research.aimultiple.com/synthetic-data-tools/ (accessed 13 September 2023)

81. Dilmegani G. The Ultimate Guide to Synthetic Data in 2023, available at: https://research.aimultiple.com/synthetic-data/ (accessed 13 September 2023)

82. Dilmegani G. Synthetic Data Generation: Techniques, Best Practices & Tools, available at: https://research.aimultiple.com/synthetic-data-generation/ (accessed 13 September 2023)

83. CHerepanov F.M., YAsnickij L.N. Laboratornyj praktikum po nejrosetevym tekhnologiyam. *Perspektivnye tekhnologii iskusstvennogo intellekta: Sbornik trudov Mezhdunarodnoj nauchno-prakticheskoj konferencii. (Penza, Penzenskij un-t, Nauchnyj Sovet RAN po metodologii iskusstvennogo intellekta, 1-6 iyulya 2008 g.) / Penz. un-t., Penza.* 2008, pp. 128–130.

84. CHerepanov F.M., YAsnickij L.N. Laboratornyj praktikum po nejrosetevym tekhnologiyam. *Svidetel'stvo o gosudarstvennoj registracii programmy dlya EVM № 2009611544. Zayavka № 2009610226. Zaregistrirovano v Reestre programm dlya EVM 12 marta 2009g.*, 2009.