ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Баранчук Дмитрий Александрович

ПРИМЕНЕНИЕ ГЛУБОКИХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ ДЛЯ ЗАДАЧ ПРОГНОЗИРОВАНИЯ В МАШИННОМ ОБУЧЕНИИ

РЕЗЮМЕ

диссертации на соискание учёной степени

кандидата компьютерных наук

Москва — 2024

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель: Бабенко Артем Валерьевич, к.ф.-м.н., Национальный исследовательский университет «Высшая школа экономики».

1 Введение

Тема диссертации

В течение последнего десятилетия возможности глубоких нейронных сетей постоянно росли, и они значительно преуспели в решении различных задач машинного обучения (ML), в таких как обработка естественного языка, распознавание изображений, синтез речи, генерация видео и многих других. Методы глубокого обучения можно разделить на два основных класса: *дискриминативные* и *генеративные* подходы.

Дискриминативные задачи пытаются ответить на конкретные вопросы касательно данных объектов. Например, определить, что изображено на картинке, подсчитать количество людей на снимках с камер видеонаблюдения или по показателям пациента предложить наиболее эффективное лечение. Более формально дискриминативные методы моделируют условное распределение $p(y|x)$ для наблюдаемых пар (x, y), где x - входной объект, а y - целевая метка. Нейронные сети быстро продемонстрировали высокую производительность в широком спектре задач прогнозирования благодаря появлению больших размеченных наборов данных и разработке специализированного оборудования, например графических процессоров (GPU). Однако в задачах распознавания все еще существует множество практических проблем. Например, в объектах данных могут быть пропущены признаки, которые могли бы быть информативными для более точных прогнозов модели. Другой пример, иногда сбор большого набора размеченных данных может быть сложной и дорогостоящей задачей, и поэтому требуются методы, которые обеспечивают максимальную точность, имея доступ только к нескольким размеченным образцам. Кроме того, данные могут подпадать под действие правовых положений GDPR и содержать личные или конфиденциальные данные пользователей. Это ограничивает использование и сбор таких данных для разработки методов машинного обучения.

В свою очередь, основная цель генеративного моделирования состоит в том, чтобы приближать распределение данных p_{data} по конечному набору наблюдаемых объектов $\mathcal{D} = \{x_0, ..., x_N\}$ из этого распределения. Текущие методы генерации аппроксимируют p_{data}, используя глубокую нейронную сеть с параметрами θ. В процессе обучения подбираются параметры, позволяющие минимизировать расстояние между распределением модели p_θ и p_{data}: $\theta^* = \min_\theta d(p_{data}, p_\theta)$. Расстояние $d(\cdot, \cdot)$ может быть произвольной мерой близости между распределениями, например, KL-дивергенция.

Наглядный пример генеративной задачи: имея картины Винсента ван Гога, обучить модель θ рисовать новые картины в том же стиле. Сравнив с похожей дискриминативной задачей "Кто нарисовал эту картину?", можно сделать вывод, что генеративные задачи обычно значительно сложнее.

Тем не менее, исследования в глубоком генеративном моделировании значительно продвинулись за последние годы. Сегодня существует множество классов глубоких генеративных моделей, которые можно разделить на две основные категории: *модели, основанные на правдоподобии* и *неявные генеративные модели*. Модели, основанные на правдоподобии, приближают p_θ, непосредственно максимизируя правдоподобие данных или его нижнюю границу. Примерами методов, основанных на принципе правдоподобия, являются авторегрессионные модели [1], диффузионные вероятностные модели [2, 3], нормализующие потоки [4, 5, 6], вариационные автокодировщики [7]. С другой стороны, неявные модели не имеют прямого доступа к функции плотности или ее оценке, но все же могут порождать правдоподобные выборки из заданного распределения. Ярким представителем являются генеративные состязательные сети (GAN) [8]. Каждый класс генеративных моделей имеет свои сильные и слабые стороны. По этой причине различные типы генеративных моделей могут быть предпочтительнее в различных практических приложениях и областях. Мы предлагает читателю однакомиться с подробным обзором существующих генеративных моделей [9, 10, 11] для получения более подробной информации.

Если еще недавно генеративные модели, в основном, рассматривались в рамках академических бенчмарков и их сложно было представить в роли рабочего инструмента на практике, то сегодня люди могут создавать реалистичные изображения по текстовому описанию [12, 13] и общаться с интеллектуальными системами, такими как GPT4 [14]. Первоочередный вопрос — можем ли мы использовать эти модели не только для развлечения, но и решать с их помощью насущные прикладные задачи. Модели для генерации изображений и видео уже создают логотипы и рекламные ролики и являются перспективным инструментом в графических редакторах. Текстовые модели активно используются для исправления ошибок и улучшения текстов, занимаются написанием качественных черновиков, тем самым уменьшая затраты на копирайтинг, или же становятся востребованными сотрудниками службы поддержки.

Однако, раз генеративные модели стали настолько “интеллектуальными”, то, возможно, имеет смысл воспользоваться ими для помощи в решении более простых дискриминативных задач? Многие исследовательские работы уже дали утвердительный ответ на этот вопрос в различных областях [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Данная диссертация расширяет это направление работ и рассматривает глубокие генеративные модели для следующих практических применений: i) заполнение пропущенных значений во временных рядах для повышения качества методов классификации и регрессии; ii) семантическая сегментация изображений при малом количестве размеченных данных; iii) генерация реалистичных и при этом приватных синтетических табличных данных для последующего решения задач машинного обучения в сценариях, когда важно не допустить утечки пользовательских данных в общий доступ.

Актуальность работы

В дипломной работе рассматриваются приложения глубоких генеративных моделей для решения трех различных фундаментальных задач машинного обучения. Ниже мы обсудим каждую из них более подробно.

Первая работа посвящена проблеме заполнения недостающих наблюдений в временных рядов, которые широко распространены в таких областях, как здравоохранение и финансы, и за последние годы их количество и сложность возросли. Недостающие значения часто возникают из-за неисправности измерительных приборов, дорогостоящих процедур и человеческих ошибок. В результате в данных могут отсутствовать некоторые информативные признаки, что приводит к тому, что методы машинного обучения делают неверные прогнозы. Недавние исследования показали, что точное заполнение пропусков во временных рядах значительно повышает производительность при выполнении последующих ML задач [26, 27, 28].

Популярные подходы на основе глубокого обучения обычно используют рекуррентные нейронные сети (RNN) для моделирования последовательностей [26, 29, 30, 28]. Другие работы объединяют RNN с состязательной задачей, например, [31, 27, 32], чтобы улучшить качество восстановленных значений. В данной работе предпринимается первая попытка использовать глубокие вероятностные генеративные модели для восстановления временных рядов. В частности, предлагается вари-

ационный автокодировщик (VAE), использующий гауссовскик процессы (GP) в латентном пространстве, и демонстрируется его эффективность на наборах данных с медицинскими показателями пациентов. В последующей работе [19] предлагается вероятностный метод заполнения пропусков, основанный на диффузионных моделях. Этот метод показывает более качественное восстановление данных и, тем самым, еще повышает эффективность ML подходов, решающие целевую задачу. Более того, привлекательным свойством вероятностных методов востановления данных является то, что они могут давать оценки неопределенности для прогнозируемых значений. Это свойство имеет решающее значение для интерпретируемости и надежности ML метода, особенно если предполагается интегрировать его в медицинские приложения.

Во второй работе исследуются генеративные модели для решения задачи семантической сегментации изображений. Семантическая сегментация — это фундаментальная задача компьютерного зрения, которая направлена на распознавание объектов изображения на уровне пикселей. В отличие от классификации изображений, где модель обычно определяет одну метку класса для всего изображения, семантическая сегментация стремится присвоить каждому пикселю свою метку. Это делает семантическую сегментацию более сложной задачей, которая на практике требует использования больших наборов размеченных данных. Однако сбор качественной разметки для большого набора изображений требует огромных человеческих усилий и денежных затрат. По этой причине методы, которые могут обеспечить высокое качество сегментации при наличии всего нескольких размеченных изображений, пользуются большим спросом [17, 18, 33].

Глубокие генеративные модели уже применялись для задачи семантической сегментации, а именно, GAN-ы [34]. Некоторые методы [35, 36, 37] используют наблюдение о том, что скрытое пространство GAN-ов содержит направление, которое позволяет генерировать синтетические изображения вместе с масками сегментации переднего плана и фона. Другие работы [17, 18] используют промежуточные представления GAN-ов на уровне пикселей для прогнозирования масок сегментации для сгенерированных изображений. Таким образом, мы можем увеличить размер выборки для обучения специализированных моделей для семантической сегментации. Эти методы демонстрируют многообещающие результаты в условиях, когда имеется ограниченное количество изображений, размеченных людьми.

Диффузионные вероятностные модели (DPM) демонстрируют самые современные технологии генерации изображений как с точки зрения качества, так и разнообразия [38, 12, 13]. Преимущества DPM успешно используются в таких задачах генерации, как раскрашивание изображений [39], повышение разрешения изображений [40, 41] и семантическое редактирование [42], где DPM-ы часто достигают более впечатляющих результатов, чем GAN-ы. Однако до сих пор не было изучено, могут ли DPM-ы эффективно применяться для распознавания изображений. В данной работе были происследованы промежуточные представления диффузионных моделей и обнаружено, что они содержат полезную семантическую информацию входного изображения на уровне пикселей. Следуя [17], предлагается новый метод семантической сегментации, использующий внутренние представления предобученной диффузионной модели. Результаты экспериментов демонстрируют его превосходство над другими подходами, которые также имеют доступ к ограниченному числу размеченных данных.

Наконец, в последней публикации, рассматривается применение диффузионных вероятностных моделей для генерации табличных данных. Табличные наборы данных обычно сильно отличаются между собой и имеют ограниченные размеры, в отличие от текстов или картинок, которые широко доступны в интернете и имеют однородную структуру. Более того, на практике табличные данные часто содержат приватную или чувствительную информацию и, следовательно, не могут быть опубликованы для общего пользования. Глубокие генеративные модели для табличной данных в основном используются для решения этой проблемы путем замены реальных пользовательских данных на синтетические. В то же время важно, чтобы синтетический данные обладал свойствами реального распределения, чтобы они были полезными для последующих приложений. В недавних работах было разработано множество методов генеративного моделирования, включая табличные VAE [43] и GAN-ы [43, 21, 22, 23, 24, 25, 44, 45, 46, 47]. Вдохновленные успехом диффузионных моделей в других областях, в работе предлагается TabDDPM — новая диффузионная модель, применимая к произвольным наборам табличных данных, и которая способна моделировать объекты с различными распределениями признаков. TabDDPM тщательно оценивается по широкому набору критериев, и мы показываем его превосходство над существующими GAN/VAE-альтернативами.

2 Основные результаты и выводы

Вклад. Основные результаты работы сформулированы ниже.

1. Предложена новая вероятностная модель: вариационный автокодировщик, использующий в своем латентном пространстве гауссовские процессы для эффективного моделирования временных рядов. Разработанная модель применяется для задачи заполнения пропусков во временных рядах из области компьютерного зрения и медицины и показано, что подход превосходит классические методы заполнения пропусков и те, что основанны на глубоком обучении. В результате, предложенный метод позволяет достигать более высокой точности в задачах предсказания на данных с пропусками и оценивать неопределенность заполненных значений, что важно для интерпретируемости ML методов.

2. Обнаружено, что современные диффузионные модели могут извлекать информативные представления изображений на уровне пикселей. Основываясь на этих знаниях, предложен новый подход для семантической сегментации, который превосходит предыдущие современные генеративные методы в сценарии ограниченного доступа к изображениям, размеченных людьми.

3. Предложена диффузионная модель для генерации табличных данных. Эта модель превосходит другие генеративные модели и может быть востребована на практике для замены приватных и конфиденциальных данных на синтетические. Это может стать шагом на пути к безопасному обмену внутренними данными компаний и разработки общедоступных и высококачественных методов прогнозирования.

Теоретическая и практическая значимость.

Предложенные методы и эмпирические результаты способствуют распространению генеративных моделей как инструмента для решения задач прогнозирования в машинном обучении. В сценариях, характеризующихся дефицитом размеченных данных, мы демонстрируем, что предварительно подготовленная диффузионная модель может служить либо эффективным механизмом для размножения данных, либо надежной дискриминационной моделью напрямую. В задаче заполнения пропущенных значений во временных рядах, мы показываем, что глубокое вероятностное моделирование является многообещающей парадигмой в приложениях здравоохранения,

где необходимо восстанавливать пропущенные измерения пациентов интерпретируемым образом. Более того, в диссертации представлен новый современный подход к генерации табличных данных, позволяющий обучать высококачественные методы машинного обучения в случаях, где важно не расскрыть личные данные пользователей или конфиденциальную информацию.

Результаты, выносимые на защиту.

1. Вероятностная генеративная модель на базе вариацинного кодировщика, которая использует гаусовские процессы в латентном пространстве для более качественного моделированния временных рядов;
2. Исследование внутренних представлений диффузионных моделей, выявляющее наличие полезной семантической информации о входных изображениях на уровне пикселей. Метод семантической сегментации, который эффективно использует представления изображений, извлеченные из предварительно обученных диффузионных моделей;
3. Метод генерации табличных данных с использованием диффузионных моделей. Исследование приложений сгенерированных данных для решения задач классификации и регрессии на табличных данных.

Личный вклад в результаты, выносимые на защиту. В первой работе автор отвечал за техническую часть статьи: разработку метода и проведение большинства экспериментов и анализа. Во второй работе автор предложил основные научные идеи, собрал наборы данных, реализовал метод, провел большинство экспериментов, анализ и написал текст. В третьей работе автор сформулировал ключевые идеи, организовал исследовательский проект, разработал план проведения экспериментов и помогал в написании статьи.

Публикации и апробация работы

Публикации повышенного уровня

1. *Vincent Fortuin*,* **Dmitry Baranchuk***, *Gunnar Rätsch, Stephan Mandt* GP-VAE: Deep probabilistic time series imputation. В материалах конференции International Conference on Artificial Intelligence and Statistics, 2020 (AISTATS 2020). Конференция ранга *A* по рейтингу CORE.

2. **Dmitry Baranchuk**, *Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, Artem Babenko* Label-Efficient Semantic Segmentation with Diffusion Models. В материалах конференции International Conference on Learning Representations, 2022 (ICLR 2022). Конференция ранга A^* по рейтингу CORE.

3. *Akim Kotelnikov,* **Dmitry Baranchuk**, *Ivan Rubachev, Artem Babenko* TabDDPM: Modelling Tabular Data with Diffusion Models. В материалах конференции International Conference on Machine Learning, 2023 (ICML 2023). Конференция ранга A^* по рейтингу CORE.

Доклады на научных семинарах

1. Семинар исследовательской группы биомедицинской информатики в ETH Zurich, Цюрих, 20 августа 2019. Тема: “Вариационные автокодировщики с гауссовскими процессами для моделирования временных рядов”.

2. Рождественский коллоквиум по компьютерному зрению. Москва, 27 декабря 2021. Тема: “Диффузионные модели для решения задачи сегментации”.

3. Семинар исследовательской группы Yandex Research, Москва, 24 июля 2022. Тема: “Применение диффузионных моделей для решения прикладных задач”.

Объем и структура работы. Диссертация содержит введение, описание выполненных исследований, заключение и тексты публикаций. Общий объем диссертации составляет 61 страниц.

3 Содержание работы

3.1 Глубокая вероятностная модель для заполнения временных рядов

В первой публикации рассматривается проблема восстановления многомерных временных рядов, т.е. заполнения в них пропущенных значений. Многомерные временные ряды состоят из множества скоррелированных одномерных временных рядов ("каналов") и требуют моделей для заполнения пропусков, которые учитывают как временные завимости внутри каждого канала, так и зависимости между каналами.

V. Fortuin*, D. Baranchuk*, G. Rätsch, S. Mandt. GP-VAE: Deep probabilistic time series imputation. AISTATS 2020

Мы обозначаем многомерный временной ряд длиной τ_T как $\mathbf{X} \in \mathbb{R}^{T \times d}$. Данные $\mathbf{x}_t = [x_{t1}, \ldots, x_{tj}, \ldots, x_{td}]^\top \in \mathbb{R}^d$ измеряется в T последовательных временных точках $\tau = [\tau_1, \ldots, \tau_T]^\top$ с $\tau_t < \tau_{t+1}$, для всех t и $\tau_1 = 0$.

Затем предполагаем, что любое количество элементов x_{tj} может отсутствовать. Таким образом, каждая точка данных может быть разделена на наблюдаемые и ненаблюдаемые признаки: $\mathbf{x}_t^o := [x_{tj} \,|\, x_{tj} \text{ наблюдалось}]$ и $\mathbf{x}_t^m := [x_{tj} \,|\, x_{tj} \text{ отсутствует}]$ с $\mathbf{x}_t^o \cup \mathbf{x}_t^m \equiv \mathbf{x}_t$ соответственно.

Заполнение пропущенных значений сводится к оценке истинных значений $\mathbf{X}^m := [\mathbf{x}_t^m]_{1:T}$ с учетом наблюдаемых данных $\mathbf{X}^o := [\mathbf{x}_t^o]_{1:T}$. Многие методы предполагают, что различные измерения во времени независимы, и в этом случае проблема вывода сводится к T отдельным задачам оценки $p(\mathbf{x}_t^m \,|\, \mathbf{x}_t^o)$. В случае временных рядов это предположение о независимости не выполняется, что приводит к более сложной задаче оценки $p(\mathbf{x}_t^m \,|\, \mathbf{x}_{1:T}^o)$.

Обзор метода. Метод использует вариационные автокодировщики (VAE), которые отображают многомерные временные ряды с пропущенными наблюдениями в латентное пространство, в котором каждое измерение полностью определено. В латентном пространстве временная динамика моделируется с помощью гауссовского процесса (GP). Поскольку многие характеристики в данных могут быть скоррелированы, латентое представление фиксирует эти корреляции и использует их для восстановления пропущенных значений. Более того, предварительная обработка данных в латентном пространстве побуждает модель встраивать данные в представление, в котором временная динамика более плавная и легко интерпретируема, чем в исходном пространстве данных. Наконец, декодер преобразует выученное латентное представление для оценки недостающих значений в исходном пространстве признаков. Схема предлагаемой модели представлена на Рис. 1. Модель включает в себя *генеративную* и *выводящую* модели, которые опишем более подробно ниже.

Генеративная модель. Сначала мы применяем GP в латентном пространстве вариационного автокодировщика. А именно, присваиваем латентную переменную $\mathbf{z}_t \in \mathbb{R}^k$ для каждого $\mathbf{x}_t$ и моделируем временные зависимости в этом латентном представлении, используя GP, $\mathbf{z}(\tau) \sim \mathcal{GP}(m_z(\cdot), k_z(\cdot, \cdot))$. Таким образом, мы разделяем задачи заполнения пропущенных значений и получения зависимостей между различными признаками объекта от моделирования динамических завимостей.

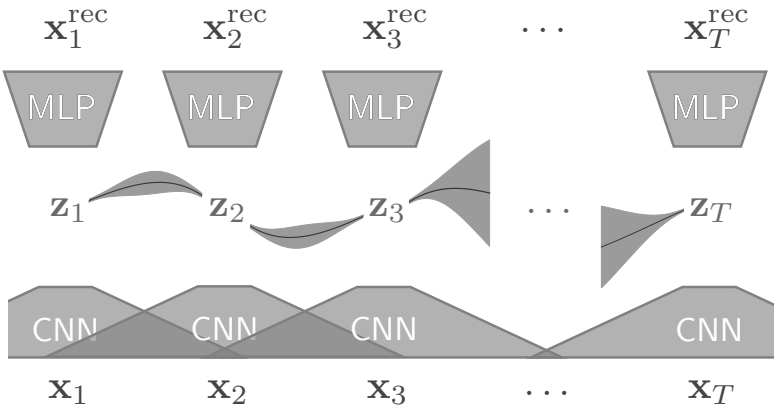


Рис. 1: Схема модели GP-VAE, состоящая из сверточного энкодера, MLP декодера и гауссовского процесса в качестве априорного распределения с функцией среднего $m(\cdot)$ и ядра $k(\cdot, \cdot)$ в латентном пространстве.

Одной из главных задач является моделирование временных рядов, которые возникают в медицинских учреждениях, где врачи измеряют различные переменные пациента и показатели его жизнедеятельности, такие как частота сердечных сокращений, кровяное давление и т.д. Практическая трудность заключается в том, что многие многомерные временные ряды в таких случаях имеют признаки с динамикой в разных временных масштабах. Для моделирования данных в различных временных масштабах мы используем ядро Коши для гауссовского процесса. Учитывая латентный временной ряд $\mathbf{z}_{1:T}$, наблюдения $\mathbf{x}_t$ генерируются по времени с помощью

$$p_\theta(\mathbf{x}_t \,|\, \mathbf{z}_t) = \mathcal{N}\left(g_\theta(\mathbf{z}_t), \sigma^2 \boldsymbol{I}\right) \ , \tag{1}$$

где $g_\theta(\cdot)$ - нелинейная функция, параметризуемая θ. В наших экспериментах функция g_θ реализовывается с помощью многослойного персептрона (MLP).

Кодировщик. Чтобы обучить параметры глубокой генеративной модели, описанной выше, и эффективно определить ее латентное пространство, нас интересует апостериорное распределение $p(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T})$. Поскольку точное апостериорное распределение трудно поддается вычислению, мы используем вариационный вывод [48, 49, 50] и амортизируем его с помощью глубокой нейронной сети [7]. Чтобы сделать вариационное распределение более выразительным и отразить временные зависимости в данных, мы используем структурированное вариационное распределение [51] с эффективным выводом, который приводит к приблизительному апостериорному распределению, которое также является GP.

Мы аппроксимируем истинное апостериорное значение $p(\mathbf{z}_{1:T,j} \mid \mathbf{x}_{1:T})$ многомерным гауссовым вариационным распределением:

$$q(\mathbf{z}_{1:T,j} \mid \mathbf{x}^o_{1:T}) = \mathcal{N}\left(\mathbf{m}_j, \mathbf{\Lambda}_j^{-1}\right) \ , \tag{2}$$

где j индексирует измерения в латентном пространстве. Такое приближение подразумевает, что вариационное апостериорное распределение может отражать зависимости во времени, но нарушает зависимости между различными измерениями в $\mathbf{z}$-пространстве (что типично для обучение VAE [7, 52]).

Мы выбираем вариационное семейство как семейство многомерных гауссовых распределений во временной области, где обратная ковариационная матрица $\mathbf{\Lambda}_j$ параметризована как трехдиагональная матрица. Таким образом, выборки из q могут быть сгенерированы за линейное время от T [53, 54, 55] в отличие от кубической временной сложности для полноранговой матрицы. Более того, по сравнению с полностью факторизованной вариационной аппроксимацией количество вариационных параметров просто удваивается. Обратите внимание, что, хотя обратная ковариационная матрица является разреженной, ковариационная матрица все еще может быть плотной, позволяя отражать долгосрочные зависимости во времени.

Мы делаем поправку в выводе по сравнению с $\mathbf{m}_j$ и $\mathbf{\Lambda}_j$, используя кодировщик $q_\psi(\cdot)$. После обучения VAE параметры генеративной модели θ и кодировщика ψ могут быть совместно обучены путем оптимизации вариационной нижней границы (ELBO),

$$\log p(\mathbf{X}^o) \geq \sum_{t=1}^{T} \mathbb{E}_{q_\psi(\mathbf{z}_t \mid \mathbf{x}_{1:T})} \left[\log \ p_\theta(\mathbf{x}^o_t \mid \mathbf{z}_t)\right] - \beta \, D_{KL}\left[q_\psi(\mathbf{z}_{1:T} \mid \mathbf{x}^o_{1:T}) \,\|\, p(\mathbf{z}_{1:T})\right] \tag{3}$$

Мы оцениваем ELBO только по наблюдаемым признакам данных, поскольку остальные признаки неизвестны, и устанавливаем для отсутствующих признаков фиксированное значение (ноль) во время применения модели.

Результаты. Проведены эксперименты на наборах данных *Healing MNIST* [58], который сочетает в себе классический MNIST [59] со свойствами, общими для медицинских временных рядов, с набором данных SPRITES [60], а также с реальным набором медицинских данных из соревнование Physionet Challenge 2012 [61]. Мы сравнили нашу модель с традиционными методами заполнения пропусков [56], методами на основе GP [62], методами на основе VAE, которые специально не предназначены для обработки временных данных [7, 57], и современными методами глубокого обучения для временных данных [63, 26] Мы наблюдаем убедительные количественные

Таблица 1: Оценка качества разных моделей для заполнения пропусков Healing MNIST и SPRITES относительно отрицательного логправдоподобия [NLL] и средней квадратичной ошибки [MSE] и также в терминах качества метода классификации, обученный на восстановленных данных [AUROC].

	Healing MNIST			SPRITES
Model	NLL	MSE	AUROC	MSE
Mean imputation [56]	-	0.168 ± 0.000	0.938 ± 0.000	0.013 ± 0.000
Forward imputation [56]	-	0.177 ± 0.000	0.935 ± 0.000	0.028 ± 0.000
VAE [7]	0.599 ± 0.002	0.232 ± 0.000	0.922 ± 0.000	0.034 ± 0.000
HI-VAE [57]	0.372 ± 0.008	0.134 ± 0.003	**0.962 ± 0.001**	0.035 ± 0.000
GP-VAE (proposed)	**0.341 ± 0.007**	**0.117 ± 0.002**	**0.960 ± 0.002**	**0.002 ± 0.000**

Таблица 2: Оценка качества моделей на данных Physionet в терминах AUROC у логистической регрессии, обученной на заполненных временных рядах.

Model	AUROC
Mean imputation [56]	0.703 ± 0.000
Forward imputation [56]	0.710 ± 0.000
GP [62]	0.704 ± 0.007
VAE [7]	0.677 ± 0.002
HI-VAE [57]	0.686 ± 0.010
GRUI-GAN [63]	0.702 ± 0.009
BRITS [26]	**0.742 ± 0.008**
GP-VAE (proposed)	**0.730 ± 0.006**

(Таблица 1, 2) и качественные (Рис. 2) подтверждения того, что GPVAE превосходит большинство базовых методов с точки зрения качества заполнения по всем трем задачам и сопоставима с передовыми моделями на медицинских данных.

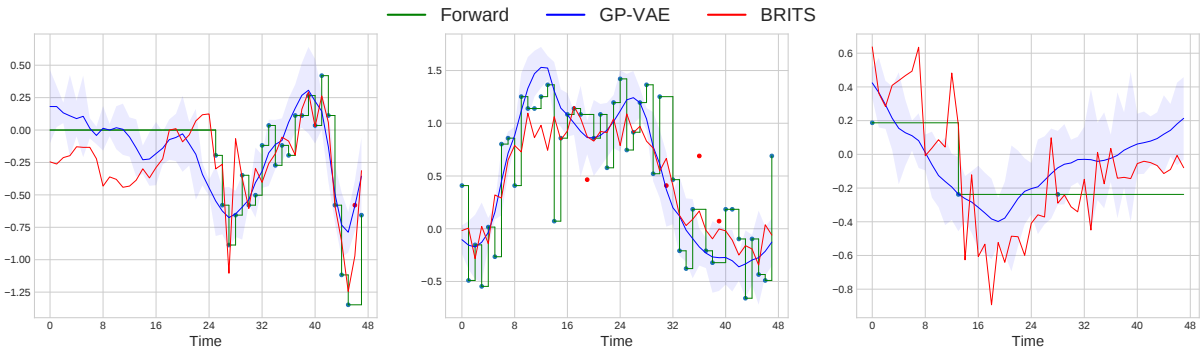


Рис. 2: Примеры заполнения пропусков в нескольких клинических объектах с разным числом изначальных пропусков. GP-VAE выдает более гладкие кривые, уменьшая шум оригинальных данных, и предлагает оценку неопределенности для предсказанных значений.

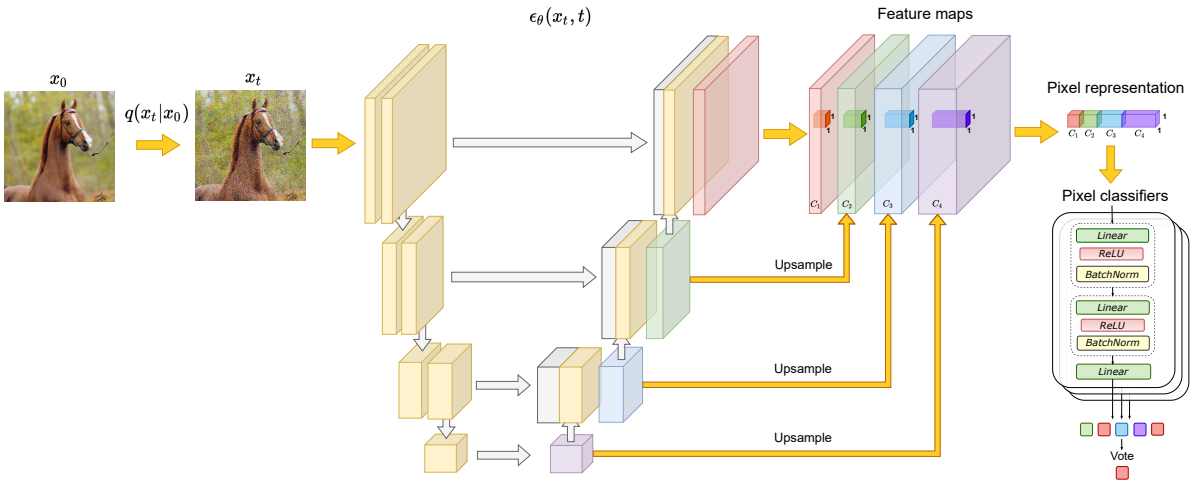


Рис. 3: **Обзор предложенного метода семантической сегментации изображений.** (1) Преобразуем изображение $x_0 \to x_t$, добавляя шум согласно $q(x_t|x_0)$. (2) Извлекаем признаки изображения из диффузинной модели $\epsilon_\theta(x_t, t)$ с разных слоев. (3) Получаем пиксельные представления, приводя извлеченные признаки к разрешению исходного изображения и объединяя их. (4) Используем полученные пиксельные представления для обучения ансамбля MLP моделей, предсказывающие метку класса для каждого пикселя.

3.2 Диффузионные модели для семантической сегментации

В следующей публикации исследуются представления, полученные с помощью современных диффузионных вероятностных моделей (DPM), и показываем, что они содержат семантическую информацию на пиксельном уровне, ценную для задачи семантической сегментации изображений.

D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, A. Babenko. Label-Efficient Semantic Segmentation with Diffusion Models. ICLR2022

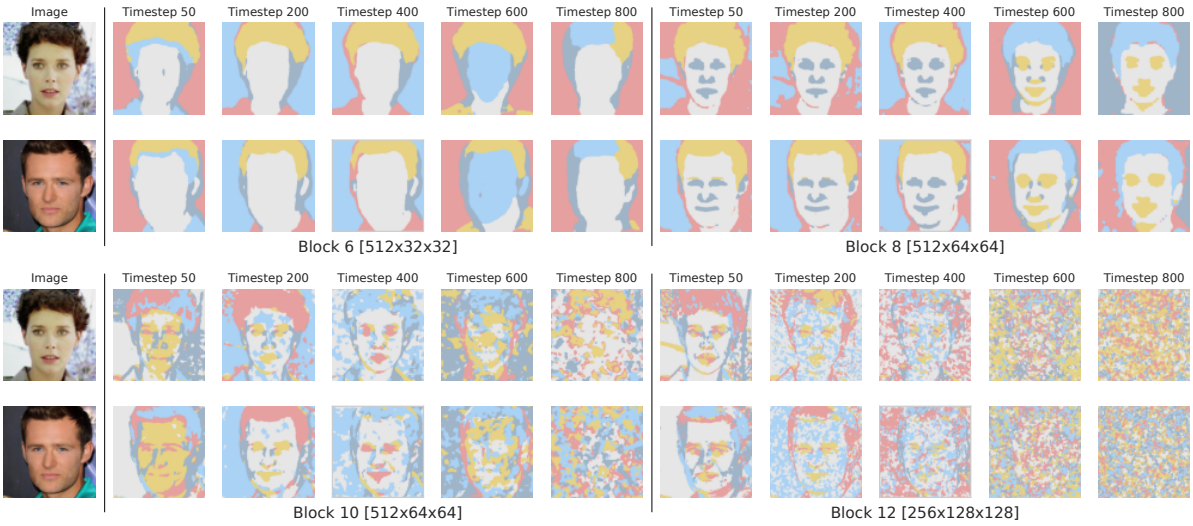


Рис. 4: Примеры k-means кластеризации (k=5) на признаках, извлеченных с разных блоков декодера UNet-а $\{6, 8, 10, 12\}$ и на разных шагах диффузии $\{50, 200, 400, 600, 800\}$. Кластеры, полученные для средних блоков, соотвествуют семантическим объектам или их частям.

Диффузионные модели. Диффузионные модели преобразуют гауссовский шум $x_T \sim N(0, I)$ в изображение x_0 путем постепенного удаления шума из x_T до менее шумных промежуточных картинок x_t. В рамках прямого диффузионного процесса зашумленный пример x_t может быть получен непосредственно из оригинальной картинки x_0:

$$\begin{aligned} q(x_t|x_0)\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \\ x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}, \ \sim \mathcal{N}(0,1), \end{aligned} \tag{4}$$

где $\alpha_t 1 - \beta_t$, $\bar{\alpha}_t \prod_{s=1}^{t} \alpha_s$ определяют расписание диффузионного процесса. Предварительно обученные DPM аппроксимируют обратный диффузионный процесс:

$$p_\theta(x_{t-1}|x_t)\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{5}$$

На практике нейронная сеть $\epsilon_\theta(x_t, t)$ предсказывает компоненту шума на временном шаге t; среднее значение представляет собой линейную комбинацию этой составляющей шума и x_t. Вместо предсказания ковариации $\Sigma_\theta(x_t, t)$ обычно используют постоянное скалярное значение для конкретного шага t.

Диффузионная модель $\epsilon_\theta(x_t, t)$ обычно параметризуется различными вариантами архитектуры UNet [64], и мы анализируем модель, предложенную в [38].

Извлечение пиксельных представлений. Для данного реального изображения $x_0 \in \mathbb{R}^{H \times W \times 3}$ можно извлечь T наборов активации из предобученной диффузинной модели $\epsilon_\theta(x_t, t)$. Общая схема для временного шага t представлена на Рис. 3.

Сначала мы зашумляем x_0, добавляя гауссовский шум в соответствии с Ур. 4. Зашумленный x_t используется в качестве входа для $\epsilon_\theta(x_t, t)$, параметризованных моделью UNet. Промежуточные активации декодера UNet-а затем преобразуются в $H \times W$ представления с помощью билинейной интерполяции. Это позволяет рассматривать их как пиксельные представления x_0.

Анализ пиксельных представлений. На Рис. 4 показаны кластеры k-means (k=5) для признаков, извлеченных из разных блоков декодера UNet-а и с разных шагов диффузии. Можно заметить, что кластеры часто соответсвуют семантическим объектам или их частям. В более глубоких блоках модели, кластера соответствуют грубым семантическим объектам. В последних блоках можно различить мелкие черты лица, но при этом признаки обладают слабой глобальной семантической информацией. Среди разных шагов диффузии наиболее информативные маски соответствуют более поздним шагам. Предполагается, что на более ранних шагах обратного процесса глобальная структура изображения у DPM еще не сформировалась, поэтому предсказать маски сегментации на таких шагах вряд ли возможно.

Метод семантической сегментации. Информативность промежуточных активаций DPM предполагает их использование для попиксильного распознавания картинок. На Рис. 3 схематично представлен предложенный подход для семантической сегментации изображений. В работе рассматривается постановка задачи, когда только для небольшого числа n обучающих изображений $\{X_1, \ldots, X_n\} \subset \mathbb{R}^{H \times W \times 3}$ есть семантические маски на K классов $\{Y_1, \ldots, Y_n\} \subset \mathbb{R}^{H \times W \times \{1,\ldots,K\}}$.

Предварительно обученная диффузионная модель используется для извлечения пиксельных представлений для размеченных изображений для подмножества блоков UNet-а и шагов диффузии t. Мы используем представления из средних блоков B=$\{5, 6, 7, 8, 12\}$ декодера UNet-а и более поздних шагов t=$\{50, 150, 250\}$ процесса обратной диффузии. Извлеченные представления из всех блоков B и шагов t приводятся к размеру изображения и объединяются, формируя векторы признаков для всех пикселей обучающих изображений. Затем, следуя [17], мы обучаем ансамбль независимых моделей многослойных персептронов (MLP), цель которых - предсказать метку класса каждого пикселя для обучающих изображений. Чтобы сегментировать тестовое изображение, DPM извлекает его попиксельные представления, и затем, они используются для прогнозирования меток всех пикселей.

Method	Bedroom-28	FFHQ-34	Cat-15	Horse-21	CelebA-19*	ADE Bedroom-30*
ALAE	20.0 ± 1.0	48.1 ± 1.3	—	—	49.7 ± 0.7	15.0 ± 0.5
VDVAE	—	57.3 ± 1.1	—	—	54.1 ± 1.0	—
GAN Inversion	13.9 ± 0.6	51.7 ± 0.8	21.4 ± 1.7	17.7 ± 0.4	51.5 ± 2.3	11.1 ± 0.2
GAN Encoder	22.4 ± 1.6	53.9 ± 1.3	32.0 ± 1.8	26.7 ± 0.7	53.9 ± 0.8	15.7 ± 0.3
SwAV	42.4 ± 1.7	56.9 ± 1.3	45.1 ± 2.1	54.0 ± 0.9	52.4 ± 1.3	30.6 ± 1.6
MAE	45.0 ± 2.0	**58.8 ± 1.1**	**52.4 ± 2.3**	63.4 ± 1.4	57.8 ± 0.4	31.7 ± 1.8
DDPM (Ours)	**49.4 ± 1.9**	**59.1 ± 1.4**	**53.7 ± 3.3**	**65.0 ± 0.8**	**59.9 ± 1.0**	**34.6 ± 1.7**

Таблица 3: Сравнение сегментационных методов согласно IoU метрике. Предложенный метод на основе DPM первосходит подходы, использующие GAN-ы и передовые self-supervised подходы.

Данные. Эксперименты проводятся на данных LSUN-Bedroom, LSUN-Cat, LSUN-Horse [65] и FFHQ-256 [34]. В качестве обучающей выборки для каждого набора данных мы рассматриваем 20-30 изображений, для которых люди размечают семантические маски. Мы обозначаем собранные наборы данных как Bedroom-28, FFHQ-34, Cat-15, Horse-21, ADE-Bedroom-30, CelebA-19, где число соответствует количеству семантических классов.

Методы. Мы сравниваем предложенный метод, основанный на DPM, с аналогичным подходами, но извлекаем признаки из других современных моделей, обученных без учителя, и генеративных моделей: MAE [66], SwAV [67], GAN Inversion + StyleGAN, GAN Encoder, VDVAE [68]

Основные результаты. Сравнение методов, используя среднее значение IoU метрики, представлено в Таблице 3. Кроме того, мы приводим несколько качественных примеров сегментации с помощью нашего метода на Рис.5. Можно заметить, что предлагаемый метод, основанный на представлениях DPM, значительно превосходит альтернативные варианты на большинстве наборах данных. Модель MAE является сильнейшим конкурентом и демонстрирует сопоставимые результаты на наборах данных FFHQ-34 и Cat-15.

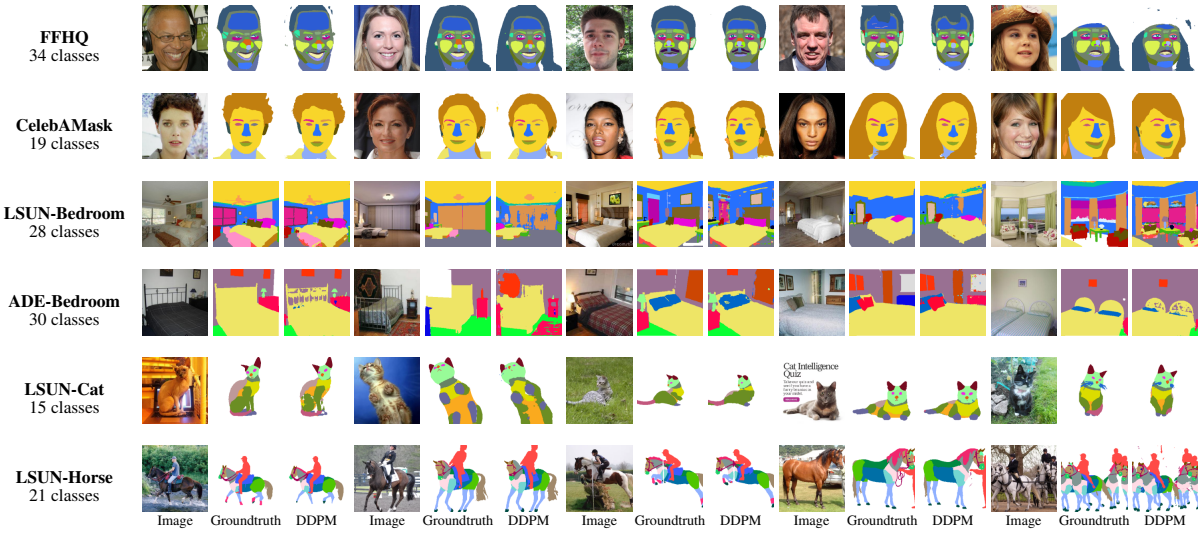
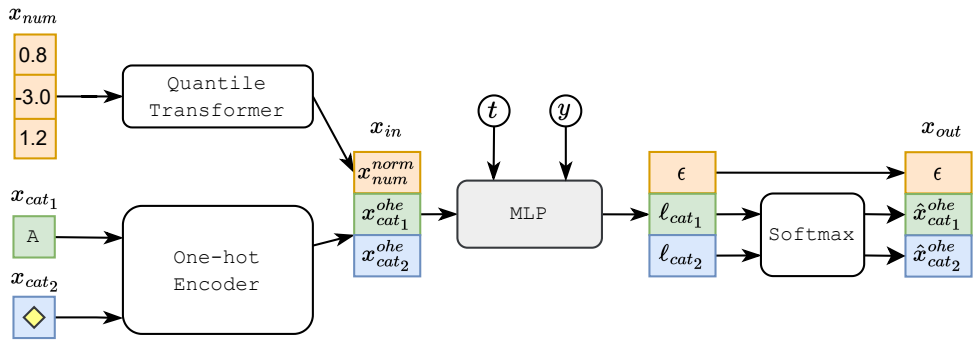


Рис. 5: Примеры масок сегментации, предсказанных нашим методом на тестовых изображениях. Рядом представлены маски, полученные людьми.

Рис. 6: Схема модели TabDDPM для классикационных табличных данных; t, y and ℓ соотвествуют шагу диффузии, метке класса, и логитам, соответственно.



3.3 Диффузионные модели для моделирования табличных данных

В рамках последней публикации, предлагается новая диффузионная модель, TabDDPM, разработанная специально для генерации табличных данных, содержащих как числовые, так и категориальные признаки. Ниже мы покажем, что предложенная модель превосходит другие генеративные подходы. Кроме того, мы увидим, что простые методы, основанные на интерполяции, такие как SMOTE [69], могут порождать удивительно реалистичные синтетические данные, обеспечивающие высокую ML эффективность. Однако в условиях конфиденциальности, когда синтети-

A. Kotelnikov, D. Baranchuk, I. Rubachev, A. Babenko. TabDDPM: Modelling Tabular Data with Diffusion Models. ICML2023

ческие данные необходимы для замены реальных пользовательских данных, которыми нельзя поделиться, предложенный метод оказывается более предпочтительное решением по сравнению со SMOTE.

Давайте разберемся в деталях предложенной модели. **TabDDPM** использует мультиномиальную диффузию для моделирования категориальных и бинарных признаков и гауссовскую диффузию для числовых признаков. В частности, табличный объект данных $x = [x_{\text{num}}, x_{\text{cat}_1}, \ldots, x_{\text{cat}_C}]$ содержит N_{num} численных признаков $x_{\text{num}} \in \mathbb{R}^{N_{\text{num}}}$ и C категориальных признаков с x_{cat_i} с K_i категориями в каждом. Что касается предварительной обработки, категориальные признаки кодируются в бинарные вектора, т.е. $x^{\text{ohe}}_{\text{cat}_i} \in \{0,1\}^{K_i}$, а числовые признаки нормализуются с использованием гауссовского квантильного преобразования из библиотеки scikit-learn [70]. Следовательно, входные данные x_0 имеют размерность $\left(N_{\text{num}} + \sum_{i=1}^{C} K_i\right)$. Шаг обратной диффузии в TabDDPM моделируется с помощью архитектуры MLP, адаптированной из [71]:

$$\begin{aligned} \texttt{MLP}(x) &= \texttt{Linear}\left(\texttt{MLPBlock}\left(\ldots\left(\texttt{MLPBlock}(x)\right)\right)\right) \\ \texttt{MLPBlock}(x) &= \texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(x))) \end{aligned} \tag{6}$$

Модель обучается путем минимизации суммы гауссовых и мультиномиальных диффузионных функций потерь. Гауссовская функция потерь соответствует L_t^{simple}[**?**] для числовых признаков. Мультиномиальная функция потерь представляет собой сумму KL дивергенций между мультиномиальными распределениями, L_t^i, для каждого категориального признака. Значения функции для мультиномиальной диффузии дополнительно делятся на количество категориальных признаков. Общая функция потерь для временного шага t может быть описана следующим образом:

$$L_t^{\text{TabDDPM}} = L_t^{\text{simple}} + \frac{\sum_{i \le C} L_t^i}{C} \tag{7}$$

Модель параметризована для прогнозирования $\epsilon \sim N(0,1)$ для числовых признаков и вероятностей категорий $\hat{x}^{\text{ohe}}_{\text{cat}_i}$ для категориальных признаков. Для наборов табличных данных для задачи классификации, модель обуславливается на метку класса, т.е. учим $p_\theta(x_{t-1}|x_t, y)$. Для регрессионных наборов данных мы рассматриваем целевое значение как дополнительную числовую характеристику и учим совместное распределение $p_\theta(x_{t-1}, y_{t-1}|x_t, y_t)$. Схематичная иллюстрация TabDDPM для классификационных наборов данных представлена на Рис. 6.

Наборы данных. Для оценки эффективности табличных генеративных моделей мы рассматриваем разнообразный набор из 15 общедоступных датасетов, ранее использовавшихся для оценки табличных моделей в [25, 71]. Эти наборы данных различаются по размеру, задаче, количеству объектов и их распределению.

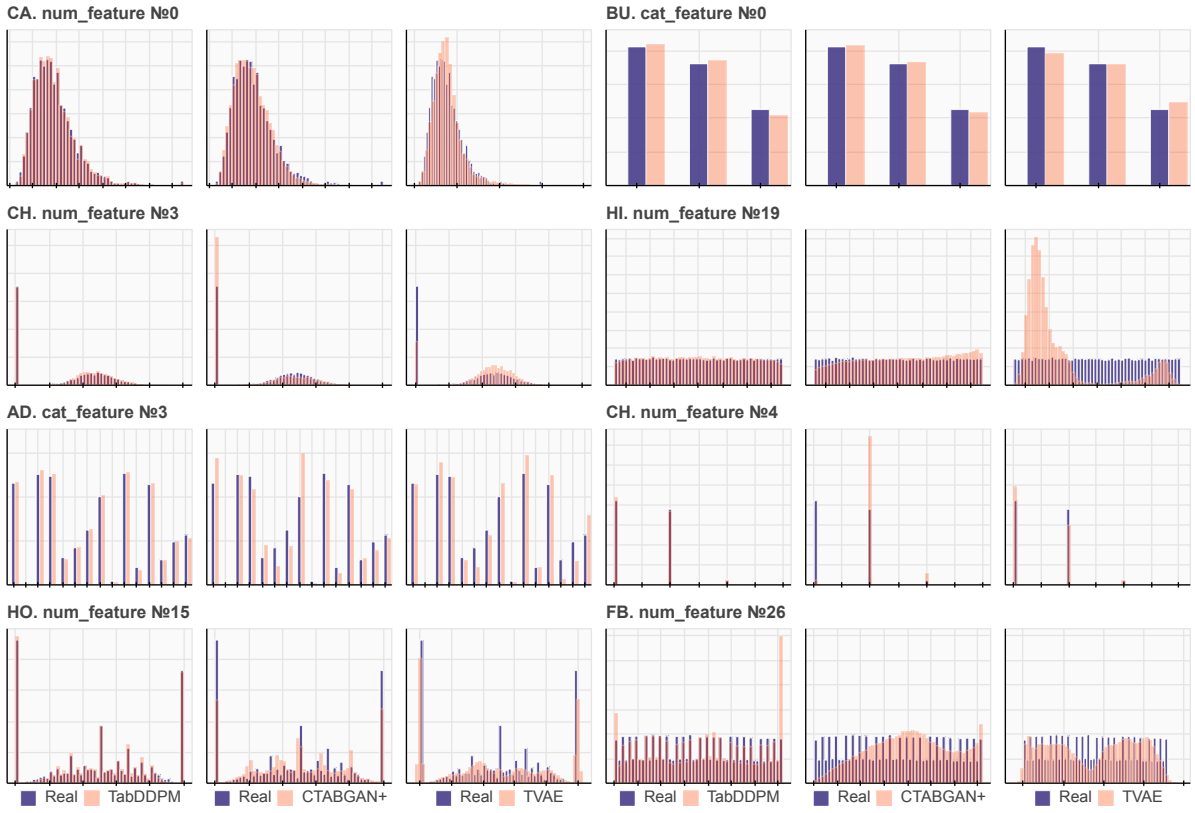
Модели для сравнения. Учитывая большое количество генеративных моделей, предложенных для табличных данных, мы оцениваем TabDDPM с ведущими подходами из каждой парадигмы генеративного моделирования: **TVAE** [43], **CTABGAN** [25], **CTABGAN+** [72]. Кроме того, мы также сравниваемся с методом, основанным на интерполяции **SMOTE** [69], и "генерируем" синтетическую точку как выпуклую комбинацию реальной точки данных и ее k-го ближайшего соседа из обучающей выборки.

Метрики качества. Нашим основным критерием оценки является ML эффективность [43]. Более подробно, ML эффективность количественно определяет качество классификационных или регрессионных моделей, обученных на синтетических данных и провалидированных на реальном тестовом наборе. В дальнейших экспериментах ML эффективность оценивается с помощью CatBoost [73], реализации GBDT, обеспечивающей самую современную производительность в табличных задачах [71]..

Качественное сравнение. Сначала мы исследуем способность TabDDPM моделировать распределение отдельных и совместных признаков. Мы визуализируем типичные распределения отдельных признаков для реальных и синтетических данных на Рис. 7.

В большинстве случаев TabDDPM показывает более реалистичные распределения по сравнению с TVAE и CTABGAN+. Преимущества более заметны (1) для числовых признаков, которые распределены равномерно, (2) для категориальных признаков с высокой кардинальностью и (3) для признаков смешанного типа, которые сочетают непрерывное и дискретное распределение. Затем мы визуализируем различия между корреляционными матрицами, вычисленными на основе реальных и синтетических данных для разных наборах признаков, см. Рис. 8. По сравнению с CTABGAN+ и TVAE, TabDDPM генерирует синтетические наборы данных с более реалистичными попарными корреляциями. Эти иллюстрации показывают, что TabDDPM является более универсальной моделью по сравнению с альтернативами и позволяет получать более качественные синтетические данные.

Рис. 7: Индивидуальные распределения признаков реальных данных и данных, сгенерированных TabDDPM, CTABGAN+ и TVAE. В большинстве случаев TabDDPM продуцирует более реалистичные распределения признаков.



ML эффективность. Затем мы сравниваем TabDDPM с альтернативными генеративными моделями с точки зрения ML эффективности. Из каждой генеративной модели мы получаем синтетический набор данных размером с оригинальный датасет. Затем эти синтетические данные используются для обучения модели классификации/регрессии. В наших экспериментах качество классификации оценивается по F1 метрике, а качество регрессии - по R2.

ML эффективность рассчитывается с использованием современных методов прогнозирования для табличных данных. В частности, для оценки используются CatBoost [73] и MLP архитектура из [71].

Основные результаты. Значения ML эффективности представлены в Таблице 4. TabDDPM значительно превосходит TVAE и CTABGAN+ на большинстве наборах данных, что подчеркивает преимущество диффузионных моделей для табличных данных. Метод, основанный на простой интерполяции, SMOTE, демонстрирует каче-

Рис. 8: Абсолютная разница между корреляционными матрицами, вычисленными на основе реальных и синтетических наборов данных. Более интенсивный красный цвет указывает на большую разницу между значениями реальной и синтетической корреляции между признаками. В большинстве случаев TabDDPM моделирует корреляции признаков лучше, чем альтернативные подходы.

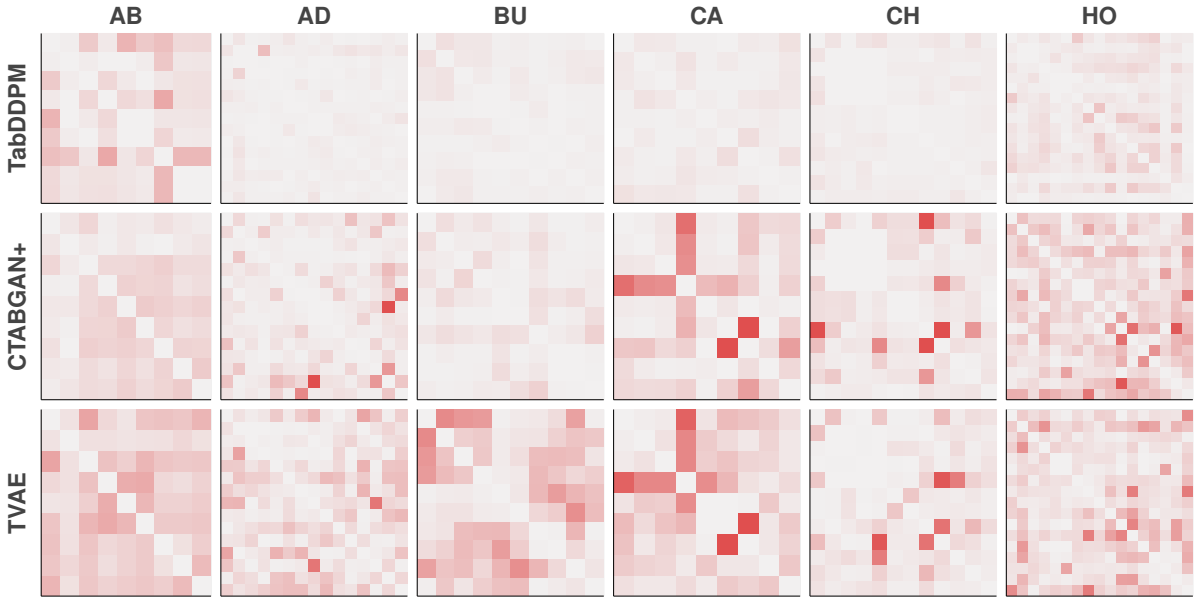


Таблица 4: Значения ML эффективности для разных генеративных моделей. Для оценки ML эффективности используется модель CatBoost.

	AB $_{(R2)}$	AD $_{(F1)}$	BU $_{(F1)}$	CA $_{(R2)}$	CAR $_{(F1)}$	CH $_{(F1)}$	DE $_{(F1)}$	DI $_{(F1)}$
CTGAN	$0.420_{\pm.004}$	$0.789_{\pm.001}$	$0.867_{\pm.003}$	$0.686_{\pm.003}$	$0.730_{\pm.001}$	$0.723_{\pm.006}$	$\mathbf{0.699_{\pm.002}}$	$0.459_{\pm.096}$
TVAE	$0.433_{\pm.008}$	$0.781_{\pm.002}$	$0.864_{\pm.005}$	$0.752_{\pm.001}$	$0.717_{\pm.001}$	$0.732_{\pm.006}$	$0.656_{\pm.007}$	$\mathbf{0.714_{\pm.039}}$
CTABGAN	–	$0.783_{\pm.002}$	$0.855_{\pm.005}$	–	$0.717_{\pm.001}$	$0.688_{\pm.006}$	$0.644_{\pm.011}$	$\mathbf{0.731_{\pm.022}}$
CTABGAN+	$0.467_{\pm.004}$	$0.772_{\pm.003}$	$0.884_{\pm.005}$	$0.525_{\pm.004}$	$0.733_{\pm.001}$	$0.702_{\pm.012}$	$0.686_{\pm.004}$	$\mathbf{0.734_{\pm.020}}$
SMOTE	$\mathbf{0.549_{\pm.005}}$	$0.791_{\pm.002}$	$0.891_{\pm.003}$	$\mathbf{0.840_{\pm.001}}$	$0.732_{\pm.001}$	$0.743_{\pm.005}$	$0.693_{\pm.003}$	$0.683_{\pm.037}$
TabDDPM	$\mathbf{0.550_{\pm.010}}$	$\mathbf{0.795_{\pm.001}}$	$\mathbf{0.906_{\pm.003}}$	$0.836_{\pm.002}$	$\mathbf{0.737_{\pm.001}}$	$\mathbf{0.755_{\pm.006}}$	$0.691_{\pm.004}$	$\mathbf{0.740_{\pm.020}}$
Real	$0.556_{\pm.004}$	$0.815_{\pm.002}$	$0.906_{\pm.002}$	$0.857_{\pm.001}$	$0.738_{\pm.001}$	$0.740_{\pm.009}$	$0.688_{\pm.003}$	$0.785_{\pm.013}$

	FB $_{(R2)}$	GE $_{(F1)}$	HI $_{(F1)}$	HO $_{(R2)}$	IN $_{(R2)}$	KI $_{(R2)}$	MI $_{(F1)}$	WI $_{(F1)}$
CTGAN	$0.443_{\pm.005}$	$0.333_{\pm.013}$	$0.575_{\pm.006}$	$0.433_{\pm.005}$	$0.745_{\pm.009}$	$0.772_{\pm.005}$	$0.783_{\pm.005}$	$0.749_{\pm.015}$
TVAE	$0.685_{\pm.003}$	$0.434_{\pm.006}$	$0.638_{\pm.003}$	$0.493_{\pm.006}$	$0.784_{\pm.010}$	$0.824_{\pm.003}$	$0.912_{\pm.001}$	$0.501_{\pm.012}$
CTABGAN	–	$0.392_{\pm.006}$	$0.575_{\pm.004}$	–	–	–	$0.889_{\pm.002}$	$\mathbf{0.906_{\pm.019}}$
CTABGAN+	$0.509_{\pm.011}$	$0.406_{\pm.009}$	$0.664_{\pm.002}$	$0.504_{\pm.005}$	$0.797_{\pm.005}$	$0.444_{\pm.014}$	$0.892_{\pm.002}$	$0.798_{\pm.021}$
SMOTE	$\mathbf{0.803_{\pm.002}}$	$\mathbf{0.658_{\pm.007}}$	$\mathbf{0.722_{\pm.001}}$	$0.662_{\pm.004}$	$\mathbf{0.812_{\pm.002}}$	$\mathbf{0.842_{\pm.004}}$	$0.932_{\pm.001}$	$\mathbf{0.913_{\pm.007}}$
TabDDPM	$0.713_{\pm.002}$	$0.597_{\pm.006}$	$\mathbf{0.722_{\pm.001}}$	$\mathbf{0.677_{\pm.010}}$	$0.809_{\pm.002}$	$\mathbf{0.833_{\pm.014}}$	$\mathbf{0.936_{\pm.001}}$	$\mathbf{0.904_{\pm.009}}$
Real	$0.837_{\pm.001}$	$0.636_{\pm.007}$	$0.724_{\pm.001}$	$0.662_{\pm.003}$	$0.814_{\pm.001}$	$0.907_{\pm.002}$	$0.934_{\pm.000}$	$0.898_{\pm.006}$

ство, сопоставимое с TabDDPM, и часто значительно превосходит ведущие подходы на основе GAN/VAE.

В итоге, TabDDPM обеспечивает передовое качество генерации и может использоваться в качестве источника реалистиных синтетических данных. Интересно, что с точки зрения ML эффективности простой метод SMOTE конкурирует с TabDDPM, что ставит вопрос о том, нужны ли сложные глубокие генеративные модели для этой задачи впринципе.

Приватность сгенерированных данных. Здесь TabDDPM рассматривается в постановках, связанных с обеспечением приватности данных, а именно, использование данных без раскрытия личной или конфиденциальной информации. В таких постановках мы стремимся создавать реалистичные синтетические данные, которые не содержат примеров из исходного датасета.

Приватность сгенерированных данных измеряется как среднее расстояние до ближайшей точки (DCR) [25] из исходного набора данных. Низкие значения DCR указывают на то, что синтетические выборки по сути копируют некоторые реальные точки и могут нарушать требования безопасности. Более высокие значения DCR указывают на то, что генеративная модель может создавать "новые"объекты, а не просто слегка искаженные дубликаты реальных данных. Обратите внимание, что данные, не имеющие никакого отношения к обучающей выборке, например, случайный шум, также обеспечивают высокий уровень DCR. Следовательно, DCR необходимо рассматривать вместе с ML эффективностью.

Таблица 5 представляет значения DCR для TabDDPM, SMOTE, CTABGAN+ и TVAE. TabDDPM более приватен, чем SMOTE, и менее приватен, чем альтернативные подход на GAN/VAE. Тем не менее, GAN/VAE-подходы обладают значительно более низкой ML эффективностью чем TabDDPM.

	AB	AD	BU	CA	CAR	CH	DE	DI
TVAE	0.088	0.220	0.226	0.056	0.010	0.241	0.096	0.146
CTABGAN+	0.081	0.400	0.242	0.070	0.020	0.235	0.131	0.204
SMOTE	0.018	0.082	0.080	0.016	0.007	0.099	0.054	0.074
TabDDPM	0.061	0.295	0.168	0.045	0.016	0.166	0.061	0.308

	FB	GE	HI	HO	IN	KI	MI	WI
TVAE	1.418	0.171	0.497	0.127	0.102	0.200	0.025	0.020
CTABGAN+	0.666	0.169	0.533	0.129	0.124	0.390	10.761	0.027
SMOTE	0.264	0.041	0.209	0.066	0.050	0.090	0.012	0.009
TabDDPM	0.785	0.076	0.473	0.096	0.050	0.252	0.574	0.023

Таблица 5: Сравнение по среднему расстоянию до ближайшей точки (DCR) (чем больше, тем лучше). TabDDPM обеспечивает лучшие показатели DCR по сравнению со SMOTE, но уступает TVAE и CTABGAN+. Мы объясняем это значительно более низкой ML эффективностью альтернатив на основе GAN/VAE.

4 Заключение

В заключительном разделе мы суммаризируем основные результаты, предложенные в диссертации:

1. Разработан подход для заполнения пропущенных значений в многомерных временных рядах с использованием новой глубокой вероятностной модели, GP-VAE, которая сочетает преимущества вариационных автоавтокодировщиков и гауссовских процессов. Модель переводит входные данные с пропусками в латентное пространство, где каждое измерение известно. Затем временные зависимости в латентном пространстве моделируются с помощью гауссовского процесса. В экспериментах показано, что предложенная модель заполняет пропуски в данных лучше чем ранее предложенные методы, и тем самым удается повысить эффективность методов прогнозирования для данных с высокой долей пропусков.

2. Проведено исследование, которое показывает, что предварительно обученные диффузионные модели могут быть эффективно использованы для извлечения семантических представлений из картинок для распознавания изображений, например для задачи семантической сегментации. Этот подход имеет ряд преимуществ по сравнению с GAN-ами: (1) более высокое качество диффузионных моделей транслируется в более информативные признаки, и (2) диффузионные модели предоставляют возможность извлекать признаки из реальных изображений напрямую. Эти преимущества позволяют диффузионным моделям обеспечивать наилучшие результаты в задаче семантической сегментации, когда есть только несколько размеченных примеров, и превосходить многие современные подходы обучения без учителя.

3. Изучен потенциал диффузионных моделей для моделирования табличных данных и предложен новый метод, TabDDPM, который генерирует данные с различными типами признаков одновременно: числовыми, порядковыми или категориальными. Показано, что TabDDPM генерирует значительно более реалистичные табличные данные, чем предыдущие подходы, основанные на GAN и VAE. В результате полученные синтетические данные могут быть использованы для обучения методов прогнозирования для задач классификации и регрессии в условиях, когда важна приватность пользовательских данных.

Список литературы

[1] Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *The Journal of Machine Learning Research*, 17(1), 2016.

[2] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[4] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf`.

[5] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *CoRR*, abs/1410.8516, 2015.

[6] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, 07–09 Jul 2015.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[9] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers, 2023.

[10] Harshvardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, 2020.

[11] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=08Yk-n5l2Al`.

[14] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[15] Anonymous. Synthetic data from diffusion models improves imagenet classification. *Submitted to Transactions on Machine Learning Research*, 2023. URL `https://openreview.net/forum?id=DlRsoxjyPm`. Under review.

[16] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. URL `https://openreview.net/forum?id=laWYA-LXlNb`.

[17] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021.

[18] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *CVPR*, 2021.

[19] YUSUKE TASHIRO, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[20] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *CVPR*, 2021.

[21] Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021.

[22] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[23] Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. Relational data synthesis using generative adversarial networks: A design space exploration. *arXiv preprint arXiv:2008.12763*, 2020.

[24] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022.

[25] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.

[26] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, pages 6775–6785, 2018.

[27] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2gan: End-to-end generative adversarial network for multivariate time series imputation. IJCAI'19, page 3094–3100. AAAI Press, 2019. ISBN 9780999241141.

[28] Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=4c0J6lwQ4_`.

[29] Jinsung Yoon, William R. Zame, and Mihaela van der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 2019.

[30] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values, 2017.

[31] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. *International Conference on Learning Representations*, 2019.

[32] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *International Conference on Machine Learning*, 2018.

[33] Nico Catalano and Matteo Matteucci. Few shot semantic segmentation: a review of methodologies and open challenges, 2023.

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[35] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. *ICML*, 2021.

[36] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020.

[37] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021.

[38] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. 2021.

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. 2021.

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. 2021.

[41] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. 2021.

[42] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. 2021.

[43] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] Jayoung Kim, Jinsung Jeon, Jaehoon Lee, Jihyeon Hyeong, and Noseong Park. Oct-gan: Neural ode-based conditional tabular gans. In *Proceedings of the Web Conference 2021*, pages 1506–1515, 2021.

[45] Yishuo Zhang, Nayyar A Zaidi, Jiahui Zhou, and Gang Li. Ganblr: a tabular data generation model. In *2021 IEEE International Conference on Data Mining (ICDM)*, 2021.

[46] Richard Nock and Mathieu Guillame-Bert. Generative trees: Adversarial and copycat. *ICML*, 2022.

[47] Bingyang Wen, Yupeng Cao, Fan Yang, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli. Causal-tgan: Modeling tabular data using causally-aware gan. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[48] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37 (2):183–233, 1999.

[49] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[50] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[51] Martin J Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 2008.

[52] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning*, 2014.

[53] Y Huang and WF McColl. Analytical inversion of general tridiagonal matrices. *Journal of Physics A: Mathematical and General*, 30(22):7919, 1997.

[54] Ranjan K Mallik. The inverse of a tridiagonal matrix. *Linear Algebra and its Applications*, 325(1-3):109–139, 2001.

[55] Robert Bamler and Stephan Mandt. Structured black box variational inference for latent time series models. *arXiv preprint arXiv:1707.01069*, 2017.

[56] Roderick JA Little and Donald B Rubin. Single imputation methods. *Statistical analysis with missing data*, pages 59–74, 2002.

[57] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

[58] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[59] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[60] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *International Conference on Machine Learning*, 2018.

[61] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

[62] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[63] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1596–1607, 2018.

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[65] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[66] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[67] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[68] Rewon Child. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

[69] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[71] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[72] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan+: Enhancing tabular data synthesis. *arXiv preprint arXiv:2204.00401*, 2022.

[73] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.