На правах рукописи

Ананьев Владислав Валерьевич

Метод и программные средства адаптивного синтеза изображений высокого разрешения на основе диффузионных моделей при работе с гигапиксельными изображениями

Специальность 2.3.5 –
“Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей”

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2026

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В. П. Иванникова Российской Академии Наук

Научный руководитель: **Аветисян Арутюн Ишханович**
доктор физико-математических наук, академик РАН

Официальные оппоненты: **Крылов Андрей Серджевич**,
доктор физико-математических наук, профессор, профессор кафедры математической физики Факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет им.М.В.Ломоносова»

Шараев Максим Геннадьевич,
кандидат физико-математических наук, руководитель лаборатории Центра прикладного ИИ Автономной некоммерческой образовательной организации высшего профессионального образования «Сколковский институт науки и технологий»

Ведущая организация: Автономная некоммерческая организация высшего образования «Университет Иннополис»

Защита состоится 17 сентября 2026 г. в 13-00 на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В.П. Иванникова Российской академии наук по адресу: 115035, г. Москва, ул. Садовническая, д. 41, ст. 2.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Института системного программирования им. В.П. Иванникова Российской академии наук.

Автореферат разослан “____” _____ 2026 г.

Ученый секретарь
диссертационного совета 24.1.120.01,
кандидат физико-математических наук Трудаков Д. Ю.

Общая характеристика работы

Актуальность темы исследования. Методы обработки изображений с применением глубоких нейронных сетей (далее ГНС), получившие активное развитие в начале 2010-х годов, привели к их успешному внедрению в широком спектре задач компьютерного зрения. Тем не менее, в процессе обучения и применения ГНС постоянно возникают три фундаментальных проблемы: высокая вычислительная сложность, потребность в наличии крупных наборов данных и крайне высокая стоимость их экспертной разметки. Для преодоления дефицита аннотированных данных традиционно применяют техники аугментации, повышающие разнообразие обучающей выборки и снижающие риск переобучения. Распространены и альтернативные стратегии обучения — обучение без учителя (unsupervised learning) и самообучение (self-supervised learning), к которым относятся методы контрастивного обучения (SimCLR, BYOL) и архитектуры на основе маскирования патчей (Masked Autoencoders).

В ряде прикладных задач потенциала таких методов достаточно. Однако существуют предметные области, где классы изображений обладают рядом особенностей: специфичная топология, экстремально высокое разрешение и строгие требования к качеству структур и объектов, присутствующих на изображении. При работе с подобными данными ограничения традиционных подходов проявляются более остро. Показательный пример — анализ биомедицинских гигапиксельных изображений.

Структура гигапиксельных изображений (например, Whole-Slide Images, размер которых может превышать $10^5 \times 10^5$ пикселей) организована в виде мультимасштабной иерархической пирамиды разрешений (рисунок 1). В рамках вычислительного анализа такие изображения декомпозируются на фрагменты (тайлы) меньшего размера на различных уровнях оптического увеличения. Обработка полученных данных осложнена проблемой выраженного ковариантного сдвига (covariate shift), который возникает из-за различий в аппаратных характеристиках, матрицах камер и алгоритмах цветопередачи сканирующего оборудования. Эти факторы порождают существенные расхождения между распределениями признаков обучающей и тестовой сред. Ситуация часто усугубляется экстремальным классовым дисбалансом, при котором целевые структуры иногда занимают менее 5% площади изображения.

Общепринятым стандартом в практике обучения моделей остается применение классических методов аугментирования — колориметрической нормализации и базовых геометрических преобразований. На репрезентативных выборках такие подходы зарекомендовали себя хорошо, выполняя функцию регуляризации и повышая обобщающую способность моделей. Однако при сильном дисбалансе классов или малом объеме выборки их эффективность резко снижается: они лишь интерполируют и пространственно трансформируют значения существующих пикселей, не порождая принципиально новых валидных

топологических конфигураций объектов. Отсюда вытекает объективная необходимость разработки специализированных методов адаптивной генерации искусственных изображений.

Для решения задачи синтеза изображений ранее активно применялись генеративно-состязательные сети (Generative Adversarial Network, GAN) и вариационные автокодировщики (Variational Autoencoder, VAE). Современный этап характеризуется переходом к вероятностным диффузионным моделям (Denoising Diffusion Probabilistic Models, DDPM), в частности к латентным диффузионным моделям (Latent Diffusion Model, LDM) на базе трансформеров (Diffusion Transformer, DiT). Такой переход открывает технологическую возможность синтеза морфологически достоверных данных в рамках заданного предметного домена.

Существенный вклад в развитие методов анализа и синтеза изображений, составляющих основу данного исследования, внесли отечественные и зарубежные научные школы компьютерного зрения, обработки изображений и генеративного моделирования. В российской научной традиции следует отметить работы школы Ю. В. Визильтера в области компьютерного зрения и распознавания визуальных объектов, работы Д. С. Ватолина и его научной школы по обработке изображений, видеоанализу и оценке качества визуальных данных, а также исследования научных коллективов ФИЦ ИУ РАН, Сколтеха и Института искусственного интеллекта AIRI, развивающих современные методы машинного обучения, мультимодального анализа данных и генеративного искусственного интеллекта. Среди зарубежных направлений ключевое значение имеют работы Я. Гудфеллоу и соавторов из Университета Монреаля по генеративно-состязательным сетям, Д. Кингмы и М. Веллинга из Амстердамского университета по вариационным автокодировщикам, а также Я. Сонга из Стэнфордского университета по вероятностным диффузионным и score-based моделям, ставшим основой современных подходов к синтезу изображений высокого разрешения.

Ключевую роль в управлении генерацией играет механизм обусловливания (conditioning), задающий характеристики синтезируемого объекта. Традиционно он опирается на текстовые описания (text prompt) либо на визуальные примеры (изображения-референсы). Однако при работе со сложноструктурированными изображениями, где требуется строгий контроль пространственного расположения объектов, текстовое обусловливание сталкивается с фундаментальными ограничениями: сбор исчерпывающей текстовой разметки крайне трудоемок, а естественный язык не позволяет однозначно описывать сложные морфологические паттерны, что приводит к непредсказуемым результатам синтеза. В связи с этим целесообразно сместить фокус на визуальное обусловливание на базе фундаментальных моделей. Предварительно обученные на десятках миллионов изображений, они позволяют извлекать векторы признаков с богатым семантическим описанием и выступают детерминированной

альтернативой, опирающейся на композицию структур в многомерном векторном пространстве, а не на их словесное описание. Разработка методов адаптивного управления визуальным обусловливанием и процессом обратной диффузии в LDM представляет собой важную научно-техническую задачу, решение которой обеспечит синтез высококачественных обучающих выборок, минимизацию доменного сдвига и повышение надежности прикладного программного обеспечения.

Целью диссертационной работы является разработка метода, алгоритмов и программных средств адаптивного синтеза изображений высокого разрешения на основе латентных диффузионных трансформеров с применением техник, направленных на компенсацию ковариантного сдвига и классового дисбаланса в обучающих выборках.

Для достижения поставленной цели сформулированы следующие **задачи исследования**:

1. Выполнить системный анализ методов генерации, формализовать схему латентной диффузии и разработать метод адаптивного синтеза, интегрирующий механизмы визуального обусловливания и архитектурную оптимизацию диффузионных трансформеров (KV-компрессию) для создания модульного и вычислительно эффективного конвейера.
2. Разработать метод доменной специализации латентных диффузионных моделей, объединяющий стратегию прогрессивного обучения для адаптации морфологии и концепцию низкоранговых адаптеров (LoRA) для тонкой параметрической настройки стиля под особенности конкретных наборов данных.
3. Разработать алгоритм локального контекстно-ориентированного встраивания объектов (inpainting) для направленного обогащения обучающих выборок, обеспечивающий пространственно-детерминированную интеграцию целевого объекта в фоновое изображение на уровне латентных представлений с процедурой автоматического колориметрического согласования.
4. Создать программно-алгоритмический комплекс для организации оптимизированных вычислительных процессов генерации (включая асинхронное кэширование и декомпозицию гигапиксельных слайдов), обосновать гибридную систему метрик и провести экспериментальную верификацию разработанных методов на данных медицинской визуализации.

Основные положения, выносимые на защиту:

1. Метод адаптивного синтеза изображений высокого разрешения, основанный на архитектуре латентных диффузионных трансформеров.
2. Новый метод доменной специализации латентных диффузионных моделей.

3. Алгоритм локального контекстно-ориентированного встраивания объектов.
4. Программно-алгоритмический комплекс адаптивного синтеза с модульной архитектурой и гибридная методология многофакторной оценки качества.

Научная новизна основных результатов исследования заключается в следующем:

1. Предложен метод адаптивного синтеза гистологических изображений высокого разрешения, новизна которого заключается в гибридном объединении архитектуры латентных диффузионных трансформеров (DiT), механизмов визуального обусловливания и алгоритмов сжатия матриц внимания (KV-компрессии). Формируемый таким сочетанием вычислительно эффективный архитектурный фундамент, в отличие от стандартных монолитных решений, является модульным: он обеспечивает строгий контроль над морфологической достоверностью генерируемых структур и позволяет независимо применять компоненты генеративного конвейера при пакетной обработке гигапиксельных изображений.
2. Предложен метод доменной специализации латентных диффузионных моделей, новизна которого заключается в двухэтапной декомпозиции процесса адаптации: применении прогрессивного обучения с фиксацией оптического увеличения для усвоения морфологии и использовании библиотеки низкоранговых адаптеров (LoRA) для изолированной коррекции стиля. В противовес доминирующей парадигме text-to-image, полный отказ от текстового обусловливания позволяет надежно удерживать геометрию структур базовым визуальным механизмом, в то время как стиль каждого конкретного набора данных кодируется отдельным компактным адаптером. Такое решение делает систему доменной адаптации масштабируемой, позволяя параметрически переключать визуальные характеристики синтеза без полного переобучения базовой модели.
3. Разработан алгоритм локального контекстно-ориентированного встраивания объектов, новизна которого определяется совмещением целевого объекта с фоном непосредственно в латентном пространстве признаков, дополненным автоматизированным гистограммным подбором фонового фрагмента и колориметрическим согласованием (в пространстве CIELAB). Отличие от существующих методов инпейнтинга заключается в достижении бесшовной интеграции объекта без использования дополнительных обучаемых архитектурных модулей и пространственной разметки.

4. Предложена гибридная методология оценки качества синтеза, новизна которой состоит в одновременном учете структурно-семантических свойств и генеративного разнообразия образцов. На основе данной методологии и созданного модульного программно-алгоритмического комплекса экспериментально доказана применимость синтетических данных для повышения качества обучения прикладных нейросетевых моделей в условиях экстремального классового дисбаланса

Практическая и теоретическая значимость работы. Предложен программно-алгоритмический комплекс для адаптивного синтеза гистологических изображений высокого разрешения, позволяющий преодолевать дефицит клинических данных и компенсировать доменный сдвиг, возникающий из-за различных факторов, возникающих в процессе подготовки биомедицинских изображений. Разработаны алгоритмы локального контекстно-ориентированного синтеза и доменной адаптации, с помощью которых экспериментально продемонстрировано повышение средних значений метрик качества и снижение их дисперсии при обучении и тестировании прикладных моделей сегментации и классификации в рамках многоэтапных диагностических конвейеров.

Предложенные решения обладают значительным потенциалом и могут служить основой для будущих исследований в области разработки методов условного синтеза медицинских изображений с применением моделей латентной диффузии.

Сформулированные в диссертации теоретические концепции и накопленный эмпирический опыт использованы при разработке программы для ЭВМ “Программный модуль подготовки гистологических изображений для задач машинного обучения” (свидетельство № 2024663138 от 04.06.2024 г. [5]). Разработанные методические и программные решения внедрены в научную и практическую деятельность 1-го патологоанатомического отделения ФГБУ “НМИЦ АГП им. В.И. Кулакова” Минздрава России при выполнении исследований, связанных с применением методов машинного обучения и компьютерного зрения для анализа гистологических изображений: справка о внедрении результатов диссертационной работы от 28.05.2026.

Апробация работы. Основные результаты диссертации были изложены на следующих конференциях:

1. III Международная научно-практическая конференция “Анализ данных в медицине” 09 июня 2023 года.
2. IV Международная научно-практическая конференция “Анализ данных в медицине” (Data Science in Medicine), 17 мая 2024 года.

3. VIII Сеченовский Международный биомедицинский саммит: научно-технологическая кооперация в медицинской отрасли (SIBS-2024), 6 ноября 2024 года.
4. Открытая конференция ИСП РАН, 11-12 декабря 2024 года.
5. VIII Петербургский медицинский инновационный форум с международным участием, 17 мая 2025 года.
6. В период выполнения диссертационного исследования автор являлся победителем конкурсного отбора 2024 года на назначение стипендии Президента Российской Федерации для аспирантов и адъюнктов, проводящих научные исследования в рамках реализации приоритетов научно-технологического развития Российской Федерации;приказ Минобрнауки России от 03.07.2024 г. № 428.

Личный вклад. Все представленные в рамках диссертационной работы результаты получены лично автором.

Публикации. Основные результаты диссертации изложены в 4 печатных изданиях. Работы [1-3] входят в международные системы цитирования Scopus и Web of Science. Работа [4] индексируется в РИНЦ. Также получено 1 свидетельство о государственной регистрации программы для ЭВМ [5].

В публикации [1] совместно с соавторами сформулирована задача и методика подсчета статистических показателей; автором выполнена разработка базы программного кода для подсчета статистических показателей и анализа результатов, а также часть работы по подготовке к публикации. В статьях [2, 3] автором выполнены постановка задачи исследования, техническая реализация моделей и их обучение, проверка и валидация результатов; автор участвовал в составлении протокола аннотирования данных и подготовке текста публикаций. В публикации [4] автором разработаны постановка задачи и планирование экспериментов, выполнены проверка и валидация результатов и подготовка текста публикации.

Работа выполнена при поддержке программы НЦМУ “Цифровой биодизайн и персонализированное здравоохранение” в части разработки программных средств анализа биомедицинских изображений и формирования воспроизводимых вычислительных конвейеров.

Объем и структура работы. Диссертация состоит из введения, 4 глав, заключения, списка литературы, списка сокращений и условных обозначений и 5 приложений. Полный объем диссертации 140 страниц текста с 9 рисунками и 9 таблицами. Список литературы содержит 74 наименования.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, а также формулируются цель, задачи, научная новизна и практическая значимость работы. Раскрывается фундаментальный характер проблемы, заключающейся в выраженном дефиците аннотированных выборок и сильном межклассовом дисбалансе, что напрямую снижает качество обучения нейросетевых моделей классификации и сегментации. Обосновывается необходимость применения современных генеративных моделей и синтетических данных для целенаправленного расширения обучающих выборок как наиболее перспективного варианта решения проблемы.

В **первой главе** диссертационного исследования проводится обзор предметной области и существующих генеративных архитектур, по итогам которого уточняется постановка задачи диссертационного исследования. Раздел 1.1 посвящен анализу предметной области, раздел 1.2 — сравнению современных генеративных архитектур, результатом которого становится обобщенная функциональная схема латентной диффузионной модели. В разделе 1.3 дополнительно рассматриваются механизмы управления генерацией и оценка вычислительной сложности, а в качестве алгоритмической основы для локального обогащения выборки обосновывается контекстное встраивание (inpainting). Сформированное базовое архитектурное решение служит фундаментом для положений, раскрываемых во второй главе.

Экспериментальная валидация в работе ориентирована на прикладную задачу цифровой патологии — детектирование пространственно-минорных целевых структур (лимфоваскулярной инвазии, ЛВИ) на гигапиксельных изображениях Whole Slide Images (WSI).

В разделе 1.1 рассматривается специфика гигапиксельных гистологических изображений WSI как объекта автоматизированного анализа. Изображение WSI описывается как многоуровневая иерархическая пирамидальная структура данных, пространственный масштаб каждого уровня которой задается физическим параметром MPP (количество микрометров на пиксель, μm/px). Обосновывается технологическая необходимость декомпозиции WSI на фрагменты фиксированного разрешения, продиктованная ограничениями видеопамяти графических ускорителей; в качестве базовых базовых параметров выбраны масштабы увеличения 20× и 40×, при которых сохраняются пространственно-текстурные признаки и геометрию объектов (рисунок 1).

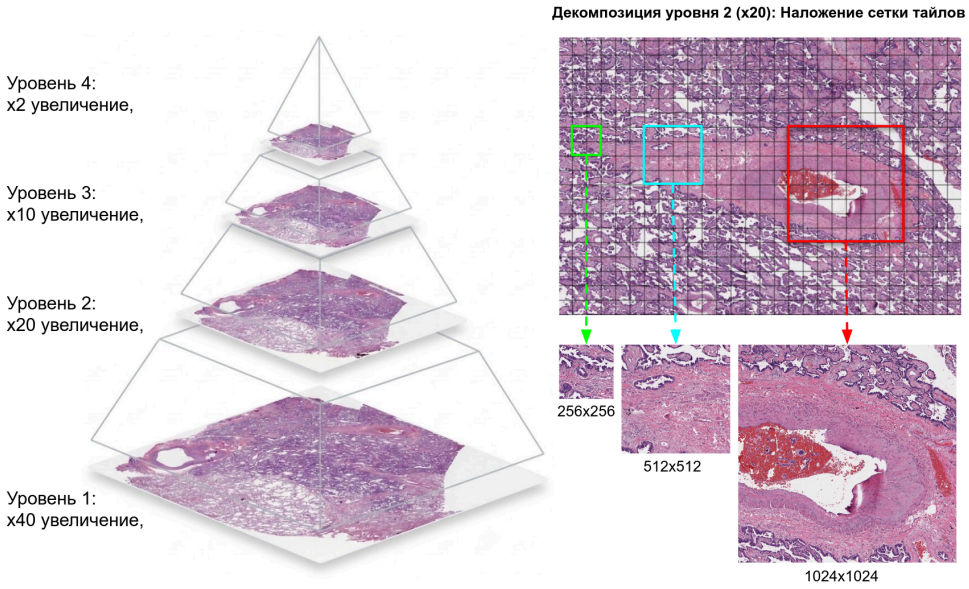

Рисунок 1 — Иерархическая структура гигапиксельного изображения WSI. Слева показана пирамида уровней оптического увеличения (×2–×40); справа приведен пример декомпозиции уровня ×20 на фрагменты различного разрешения (256×256, 512×512, 1024×1024)

Анализируется проблема доменного сдвига возникающего между средами оцифровки, частным проявлением которого выступает ковариантный сдвиг распределения признаков. В его основе — вариативность технических параметров лабораторного оборудования: спектральных характеристик источников света сканирующих систем, профилей цветопередачи матриц, алгоритмов автофокусировки и калибровки баланса белого. Как следствие, совместное распределение признаков обучающей выборки P_{train} существенно отличается от распределения в целевой тестовой среде P_{test}. Ситуация усугубляется экстремальным дисбалансом классов: из 8212 размеченных структур сосудистого русла на 207 слайдах WSI лишь 216 объектов (менее 3%) содержали признаки целевого минорного класса. Обосновывается, что классические методы аугментирования (цветокоррекция и геометрические трансформации) не компенсируют семантический дефицит выборки, поскольку лишь преобразуют существующие пиксели, не порождая новых валидных топологических конфигураций структур.

В разделе 1.2 проведен сравнительный анализ генеративных архитектур применительно к задаче параметрического синтеза изображений. Проводится обзор процесса эволюции генеративных моделей и выявляются ограничения таких подходов как, генеративно-состязательные сетеи (GAN) и вариационные автокодировщики (VAE), которые существенно усложняют и замедляют процесс обучения моделей и в значительной мере влияют на качество получаемого результата. Обосновывается, что современный этап развития ГНС характеризуется переходом к латентным диффузионным моделям (LDM) и диффузионным трансформерам (DiT), превосходящим аналоги по стабильности обучения и качеству условного

синтеза. Аргументирован концептуальный отказ от текстового управления генерацией (text-to-image) в пользу визуального обусловливания (image-to-image): естественный язык не позволяет однозначно и стабильно описывать строго детерминированную геометрию и текстуру биомедицинских изображений. Для реализации визуального контроля обосновано применение специализированных фундаментальных моделей-энкодеров (UNI, Virchow-2 и др.). Компактные векторы признаков, извлекаемые моделью UNI2-h, служат управляющим условием направленного синтеза в обход семантической неопределенности текстовых описаний.

Рассмотренные компоненты сводятся в обобщенную функциональную схему латентной диффузионной модели (рисунок 2) — базовое архитектурное решение, инвариантное к программным реализациям. Схема объединяет три компонента: подсистему перцептивного кодирования и декодирования изображений в латентное признаковое пространство на основе автокодировщика SD3-VAE; центральное генеративное ядро — диффузионный трансформер, выполняющий обратный диффузионный процесс непосредственно в латентном пространстве; модуль визуального обусловливания, направляющий генерацию эмбеддингами фундаментальной модели UNI2-h. Именно эта конфигурация, реализованная на базе архитектуры PixCell, принимается за основу разрабатываемого метода и формализуется во второй главе.

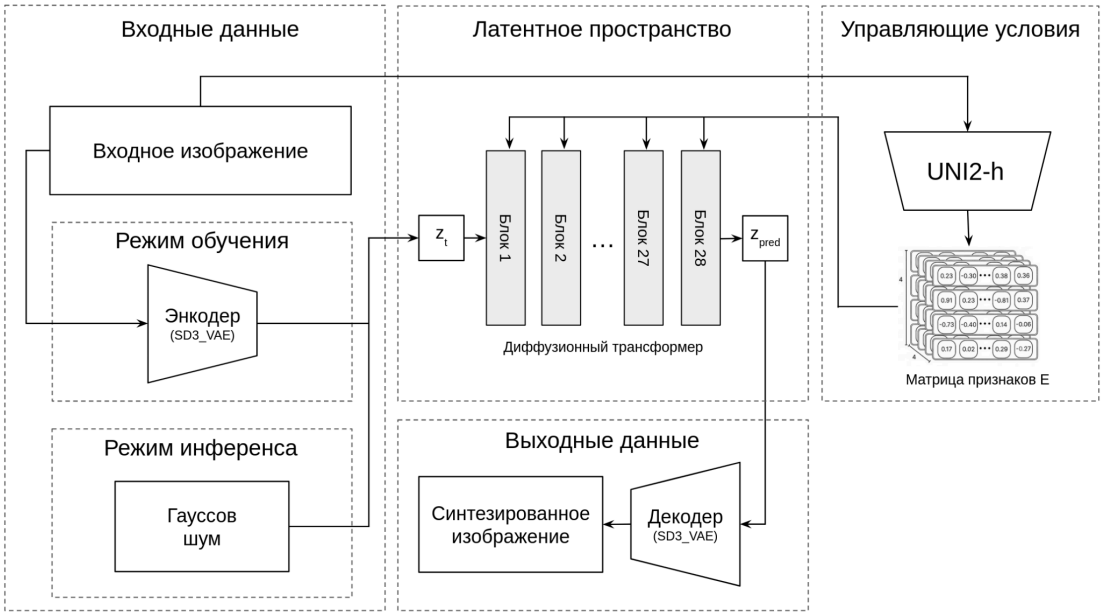

Рисунок 2 — Функциональная схема латентной диффузионной модели, демонстрирующая взаимодействие базовых компонентов (автоэнкодера и диффузионного трансформера) с механизмом визуального обусловливания на основе матрицы признаков UNI2-h в режимах обучения и инференса

В разделе 1.3 рассматриваются механизмы управления генерацией и выполняется оценка вычислительной сложности диффузионных моделей. Проанализированы специализированные

архитектурные надстройки (ControlNet, IP-Adapter); показано, что для локального обогащения выборки объектами минорного класса более адекватным алгоритмическим решением выступает контекстное достраивание (inpainting), не требующее обучения дополнительных модулей и труднодоступных парных {изображение; маска} данных. Анализ временной сложности выявил два независимых направления оптимизации: сокращение числа шагов сэмплирования и снижение квадратичной сложности механизма self-attention.

Таким образом, первая глава дает обзор предметной области и существующих генеративных решений, итогом которого становятся обобщенная функциональная схема латентной диффузионной модели и уточненная постановка задачи исследования. На этом базовом архитектурном решении строятся положения, раскрываемые во второй главе.

Во **второй главе** раскрывается концептуальное ядро исследования. На рассмотрение выносятся три положения, выносимые на защиту: метод адаптивного синтеза изображений высокого разрешения (раздел 2.1), метод доменной специализации латентных диффузионных моделей (раздел 2.2) и алгоритм локального контекстно-ориентированного встраивания объектов (раздел 2.3). Изложение опирается на принятую в первой главе базовую модель PixCell и развивает обобщенную функциональную схему до уровня математически формализованного алгоритмического конвейера.

Метод адаптивного синтеза изображений высокого разрешения

В разделе 2.1 представлена математическая модель диффузионного процесса в латентном пространстве на базе архитектуры диффузионных трансформеров (DiT) — первое положение, выносимое на защиту. Формализован переход к латентному признаковому пространству $Z = E_{VAE}(x)$ с использованием вариационного автокодировщика SD3-VAE, минимизирующего пространственные искажения и потери резкости на границах микроструктурных элементов, геометрических контуров и текстурных паттернов. Управление процессом обратной диффузии осуществляется с помощью векторов визуального обусловливания, извлекаемых фундаментальной моделью-энкодером UNI2-h, что воплощает концепцию контролируемого image conditioning и полностью исключает семантическую неопределенность текстовых описаний (подраздел 2.1.1).

В качестве ядра вычислительной оптимизации метода (подраздел 2.1.2) служит алгоритм пространственного сжатия матриц ключей и значений (KV-compression), устраняющий квадратичный рост вычислительной сложности $O(N^2)$ слоев self-attention при обработке изображений высокого разрешения (1024×1024 пикселей и выше). Внедрение механизма снижает сложность слоев внимания до $O(N^2/R^2)$, где R — коэффициент пространственного сжатия (размер стороны локального окна).

Метод доменной специализации латентных диффузионных моделей

В разделе 2.2 формулируется метод доменной специализации латентных диффузионных моделей — второе положение, выносимое на защиту. Метод решает задачу параметрического переноса стиля целевого домена без деструктивного вмешательства в базовые геометрические представления модели и реализуется в два последовательных этапа.

Первый этап — формирование набора базовых моделей средствами прогрессивного обучения (подраздел 2.2.1): модели обучаются поэтапно, с последовательным увеличением разрешения генерируемых изображений от 256×256 до 1024×1024 пикселей при строгой фиксации уровня оптического увеличения. Подход значительно ускоряет обучение, обеспечивает устойчивую сходимость функции потерь и формирует готовый набор базовых моделей W_0, адаптированных для различных разрешений. Второй этап — доменная адаптация средствами низкоранговых адаптеров (LoRA, подраздел 2.2.2): базовые модели настраиваются под визуальные характеристики конкретного набора данных путем обучения компактных матриц в слоях cross-attention при замороженных основных весах. Такое решение исключает риск “катастрофического забывания” признаков и позволяет быстро переключаться между графическими стилями доменов.

Алгоритм локального контекстно-ориентированного встраивания объектов

В разделе 2.3 описан алгоритм локального контекстно-ориентированного встраивания объектов — третье положение, выносимое на защиту. Алгоритм направлен на компенсацию дисбаланса минорных классов и повышение их генеративного разнообразия за счет совмещения структур в латентном пространстве признаков (рисунок 3). Процедура включает три ключевых шага: первичное пространственно-колориметрическое совмещение донорского объекта и фонового фрагмента в предварительный композит; генеративное «спаивание» зоны их стыка при прохождении через процессы прямой и обратной диффузии с применением управляющего семантического вектора, полученного на базе предварительного композита; финальную локализованную сборку изображения с морфологической коррекцией, возвращающей детали, которые неизбежно утрачиваются в ходе диффузионной обработки. Ограничение глубины диффузии и применение сглаженных карт влияния дают возможность сконцентрировать генеративное воздействие исключительно на границах совмещения. Такой подход позволяет сохранить исходную структуру донора и фона, обеспечивая бесшовную интеграцию (blending) объектов без использования дополнительных обучаемых модулей и возникновения визуальных артефактов, характерных для классических методов инпейнтинга.

Рисунок 3 — Концептуальная схема работы алгоритма контекстно-адаптивного встраивания объектов

Третья глава посвящена программной реализации разработанных методов и относится к четвертому положению, выносимому на защиту, в части программно-алгоритмического комплекса. Приводится описание модульной архитектуры, алгоритмического обеспечения прикладных сценариев синтеза и методов аппаратно-программной оптимизации вычислений при обработке гигапиксельных изображений на базе экосистемы фреймворка Hugging Face.

В разделе 3.1 описана модульная программная архитектура генеративного комплекса, реализующая масштабируемое взаимодействие четырех независимых функциональных контуров (рисунок 4): контура ввода-вывода (асинхронная декомпозиция гигапиксельных изображений на тайлы фиксированного разрешения средствами потокового чтения); контура латентно-признакового кодирования (вариационный автокодировщик SD3-VAE и фундаментальная модель UNI2-h для параллельного отображения в латентное признаковое пространство); генеративного контура (модифицированная магистраль DiT и динамическая библиотека адаптеров LoRA); контура прикладных задач (сопряжение с библиотекой huggeron для преобразования результатов синтеза в наборы данных, совместимые с конвейерами обучения моделей классификации и сегментации). Фиксированный формат интерфейсов обеспечивает слабую связанность компонентов и асинхронное выполнение этапов конвейера. Ядро системы реализовано на основе модификации абстракции `DiffusionPipeline` библиотеки diffusers, а динамическое распределение нагрузки на графические ускорители — модулем `DistributedDataParallel`.

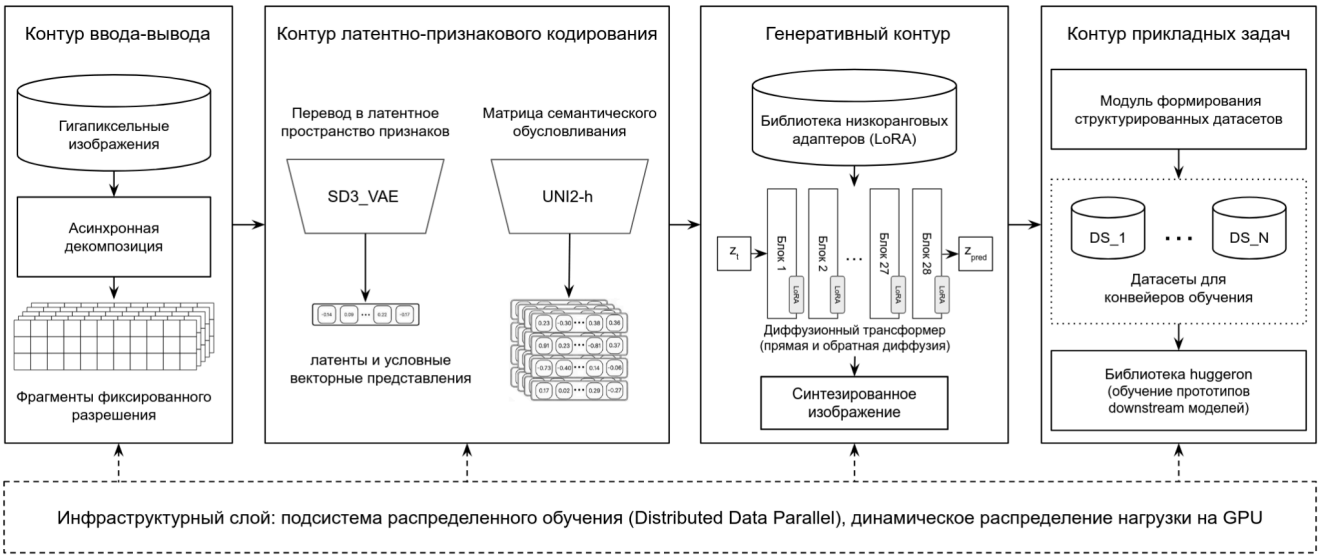

Рисунок 4 — Схема взаимодействия независимых функциональных контуров в модульной архитектуре генеративной системы

В разделе 3.2 представлено алгоритмическое обеспечение двух прикладных сценариев целенаправленного расширения обучающих выборок — доменной специализации (подраздел 3.2.1) и локального контекстно-ориентированного встраивания объектов (подраздел 3.2.2). Оба сценария переводят теоретические концепции второй главы в детерминированные вычислительные конвейеры и реализованы в виде независимых программных модулей единого комплекса. Общая библиотека конфигураций и унифицированный формат обмена данными позволяют выполнять эти этапы как в строгой последовательности, так и встраивать их во внешние системы без модификации базовых алгоритмов.

В подразделе 3.2.1 описана алгоритмическая реализация метода доменной специализации латентных диффузионных моделей, объединяющая фазы адаптации базовой архитектуры и генеративного вывода (рисунок 5). На этапе формирования моделей (левая ветвь) реализуется итеративный процесс прогрессивного обучения с повышением разрешения: для масштабов, превышающих стартовые 256 пикселей, осуществляется последовательный перенос весов с предыдущих этапов вплоть до достижения целевой размерности 1024. Полученный набор базовых моделей W_0 служит фундаментом для создания библиотеки предметно-ориентированных LoRA-адаптеров, дообучаемых точечно на слоях cross-attention. На этапе инференса (правая ветвь) модель из списка W_0 инициализируется с интегрированным LoRA адаптером из сформированной библиотеки. В процессе генерации, референсное изображение X_{ref} подается на вход DiT, в трансформерных блоках которого разворачиваются два параллельных управляющих потока: пространственное обусловливание, где извлеченная моделью UNI-2h матрица признаков $e = E_{UNI}(X_{ref}) \in \mathbb{R}^{16x1536}$ подается на KV-матриц для

строгого удержания морфологии, а дополнительная коррекция стиля осуществляется посредством контролируемой активации весов LoRA-адаптера.

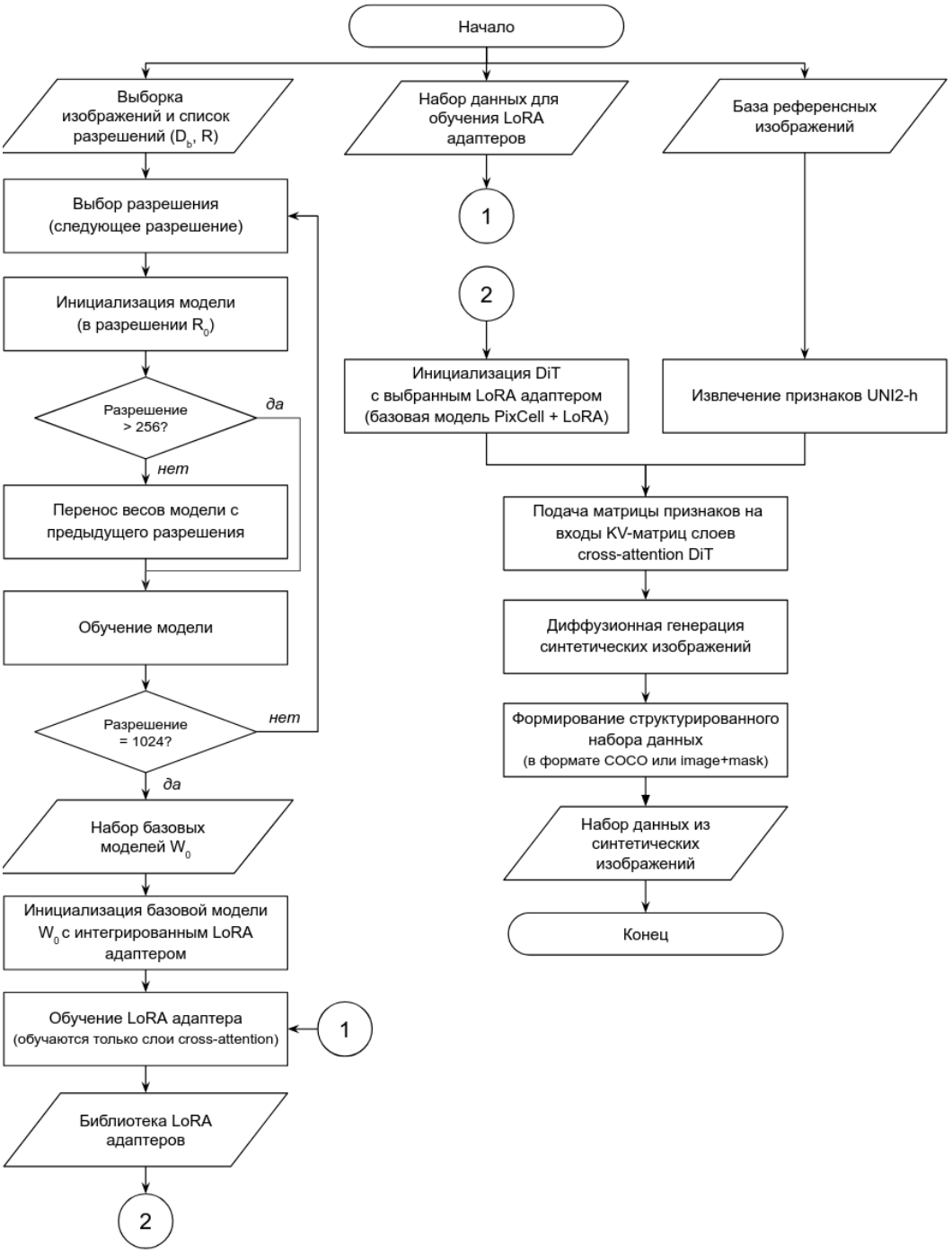

Рисунок 5 — Схема алгоритма гибкого вариативного синтеза. Левая ветвь иллюстрирует процесс дообучения и формирование библиотеки адаптеров; правая ветвь — конвейер диффузионной генерации

В подразделе 3.2.2 представлена программная реализация алгоритма локального контекстно-ориентированного встраивания объектов (рисунок 6); псевдокод приведен в листинге 1. Алгоритм поддерживает два режима подготовки фонового изображения T — использование реального фрагмента ($T_{real} \sim p_{real}(x)$) или генерацию синтетического фона (

$T_{syn} \sim p_{\theta}(x \mid E_{UNI})$) — и независимо от режима декомпозируется на три последовательных этапа.

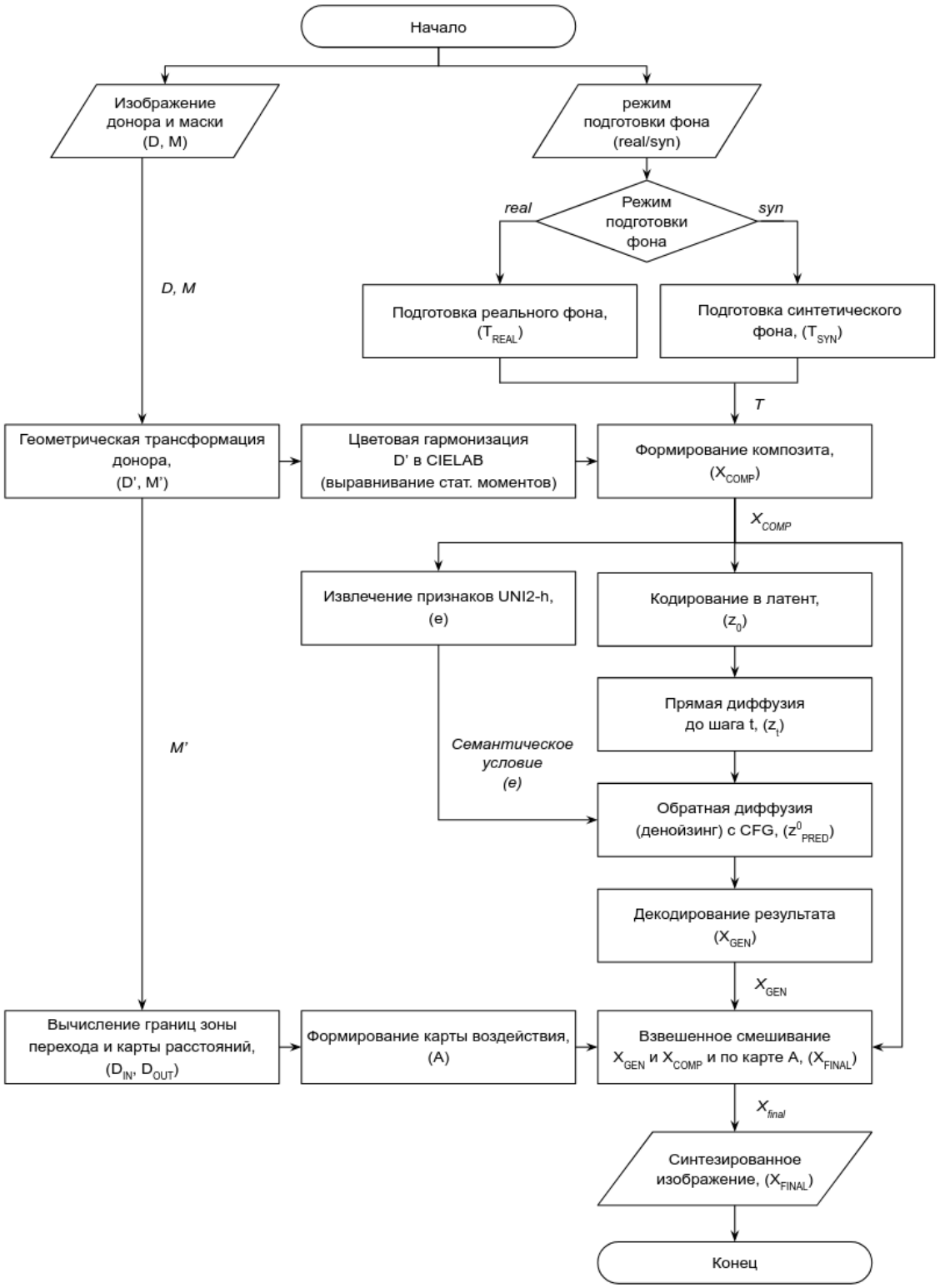

Рисунок 6 — Схема алгоритма контекстно-ориентированного inpainting на основе семантически обусловленной диффузионной модели и метода морфологического пограничного смешивания

Пространственное и хроматическое выравнивание. Вычисляются оптимальные координаты и угол поворота, трансформирующие исходного донора D и его маску в пространственно согласованные версии D′ и M′. Для устранения колориметрического рассогласования изображений донора и фона выполняется цветовая гармонизация D′ методом выравнивания статистических моментов в пространстве CIELAB. На основе бинарной маски M' формируется предварительный композит из донора D' и фона T:

$$X_{comp} = (1 - M') \odot T + M' \odot D' \quad (1)$$

Семантически обусловленная диффузионная генерация. Из композита извлекается матрица признаков $e = E_{UNI}(X_{comp}) \in \mathbb{R}^{16x1536}$. Композит переводится в латентное представление $z_0 = E_{vae}(X_{comp})$ и подвергается зашумлению в рамках прямого процесса диффузии до промежуточного временного шага t, за счет чего формируется его зашумленная версия z_t:

$$z_t = \sqrt{\bar{\alpha}_t} \cdot z_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \varepsilon, \varepsilon \sim N(0, I) \quad (2)$$

Инициализация процесса обратной диффузии из состояния z_t ограничивает глубину преобразования и сохраняет общую геометрию композита, направляя генеративное воздействие преимущественно на зону стыка фрагментов. Процесс денойзинга с контролем степени следования семантическому условию CFG (Classifier-Free Guidance) завершается декодированием D_{VAE} латентного представления z^0_{pred} обратно в пространство изображений:

$$X_{gen} = D_{vae}(z^0_{pred}) \quad (3)$$

Морфологическая коррекция границ. Вычисляются поля расстояний до границы маски изнутри и снаружи: $d_{in} = DT(M')$, $d_{out} = DT(1 - M')$, где DT(·) – преобразование расстояния. На их основе формируется карта воздействия $A = \exp(-d^2_{out}/2\sigma^2) \cdot \exp(-d^2_{in}/2\sigma^2)$ с последующей нормировкой $A \leftarrow A / max(A)$, σ = 15. Значения карты воздействия максимальны на границе и экспоненциально затухают вглубь как донора, так и фона, что позволяет сконцентрировать воздействие диффузионной модели в узкой зоне вдоль шва. Итоговое изображение вычисляется как взвешенная комбинация композита и результата диффузии:

$$X_{final} = (1 - A) \odot X_{comp} + A \odot X_{gen} \quad (4)$$

Полная математическая модель метода сворачивается в единую функцию граничного смешивания $\mathcal{B}$:

$$X_{final} = \mathcal{B}(G_\theta(E_{vae}(X_{comp}), E_{UNI}(X_{comp})), X_{comp}, M') \quad (5)$$

Листинг 1. Псевдокод алгоритма структурно-ориентированной гистологической диффузии.

Входные данные:
- T — фон;
- D — объект-донор;
- M — бинарная маска объекта.

Параметры:
- $\tau = 0{,}55$
- $\lambda = 0{,}85$
- $s = 2{,}5$
- $\sigma = 15$

Выходные данные:
- X_{FINAL} — изображение со встроенным объектом;
- M' — преобразованная маска объекта.

```
# Этап 1: Пространственное и хроматическое выравнивание
φ_T ← ОценкаОриентации(T);
φ_D ← ОценкаОриентации(D, M)
(D', M') ← ГеометрическоеПреобразование(D, M, φ_T, φ_D)
D'       ← ХроматическоеВыравнивание(D', T, M', λ)  # CIELAB
X_COMP   ← (1 − M') ⊙ T + M' ⊙ D'                    # Формирование композита

# Этап 2: Ограниченная диффузионная генерация
e   ← E_UNI(X_COMP)
z_0 ← E_VAE(X_COMP)
t   ← ⌊τ · t_MAX⌋
z_t ← √ᾱ_t · z_0 + √(1 − ᾱ_t) · ε                   # Зашумление композита
z_0^pred ← ОбратнаяДиффузия(z_t, e, s, t)           # Обратная диффузия
X_GEN ← D_VAE(z_0^pred)

# Этап 3: Морфологическое смешивание
d_IN  ← ПреобразованиеРасстояний(M')
d_OUT ← ПреобразованиеРасстояний(1 − M')
A     ← ВычислениеВесовойКарты(d_IN , d_OUT, σ)
X_FINAL ← (1 − A) ⊙ X_COMP + A ⊙ X_GEN              # Локальное слияние
вернуть X_FINAL, M'
```

В разделе 3.3 описаны методы аппаратно-программной оптимизации комплекса. Для устранения ограничений пропускной способности шины ввода-вывода при многократном чтении изображений реализована стратегия предварительного вычисления: векторы признаков UNI2-h и латентные представления SD3-VAE заблаговременно извлекаются и сохраняются в ОЗУ. Переход к агрегированному хранению и многопоточному RAM-кэшированию признаков минимизировал задержки обращения к дисковой подсистеме и время простоя GPU. Интерфейсы сопряжения с библиотекой huggeron автоматизируют формирование аннотированных наборов данных на базе результатов синтеза, замыкая единый технологический цикл систем интеллектуального анализа изображений высокого разрешения.

Четвертая глава завершает обоснование четвертого положения, выносимого на защиту, в части гибридной методологии оценки качества и экспериментального подтверждения эффективности синтетических данных. Верификация проводится по трем ключевым направлениям: вычислительная эффективность оптимизаций, морфологическая достоверность

синтезированных изображений и влияние синтетических данных на качество обучения прикладных нейросетевых моделей компьютерного зрения.

В разделе 4.1 описана методология экспериментов. Вычислительный конвейер развернут на графическом ускорителе Nvidia A100; разработка велась с использованием актуальных версий библиотек CUDA и PyTorch. Тестирование проводилось на трех выборках WSI с выраженным дефицитом аннотированных данных (таблица 1). Базовая генеративная модель PixCell предварительно обучена на многодоменном корпусе из 69184 WSI, охватывающем 28 текстурно-графических доменов, что исключает ресурсоемкое обучение весов с нуля. Разделение на обучающую и контрольную (hold-out) выборки выполнено на уровне слайдов для исключения утечки признаков; оценка прикладных моделей проводилась по схеме 10-кратной кросс-валидации.

Таблица 1 — Состав исходных клинических выборок.

Выборка	**Кол-во аннотированных WSI**	**Общее кол-во WSI**
DHMC	102	143
LCNOV	54	161
NLST	51	145
Всего	207	449

Объемы выборок реальных и синтетических изображений, подготовленных для обучения моделей на всех этапах, приведены в таблице 2. Для сегментации использованы 8212 масок кровеносных сосудов, размеченных на 207 WSI (5764 изображения); для классификации инвазии отобрано 1875 изображений двух классов («инвазия» и «чистый сосуд»). Синтетические наборы получены с применением обоих предложенных алгоритмов генерации.

Таблица 2 — Объем обучающих выборок по задачам

Этап	**Задача**	**Реальные изображения**	**Синтетические изображения**
Обучение генеративной модели	Дообучение: 256	112921	-
	Дообучение: 512	28228	-
	Дообучение: 1024	7058	-
	Обучение LoRA	10000	-
Обучение прикладных моделей	Сегментация сосудов	5764	5172
	Классификация инвазии	1875	4520

Для многофакторной оценки качества синтеза сформирована система метрик, охватывающая четыре взаимодополняющих аспекта: согласованность распределений, сохранение доменно-специфичной семантики, локальное разнообразие и уникальность образцов. Оценка проводится в признаковых пространствах нескольких моделей-кодировщиков — универсального Inception-v3 и доменно-специфичных Virchow-2 и UNI2-h. Близость распределений измеряется расстояниями Фреше и Кернела (FID и KID; KID несмещен на выборках ограниченного объема); метрика CLIP-FID чувствительна к согласованности макроструктур; пара Precision/Recall характеризует реализм и разнообразие, причем снижение Recall сигнализировало бы о “коллапсе мод”; косинусное сходство эмбеддингов оценивает сохранение семантики референса, а метрика NND (Nearest Neighbour Distance) — уникальность образцов относительно реальной выборки. Качество прикладных моделей сегментации (UNet++) и классификации (EfficientNet-B1) оценивается стандартными метриками компьютерного зрения (IoU, F1-Score, Sensitivity, Specificity), а вычислительная эффективность — пропускной способностью, временем эпохи и пиковым потреблением видеопамяти.

В разделе 4.2 представлены результаты оценки методов аппаратно-программной оптимизации (таблица 3). Предварительное вычисление латентных векторов SD3-VAE и эмбеддингов UNI2-h сократило время обучения одной эпохи на 19,5% при росте пропускной способности конвейера в 1,24 раза и кратном снижении разброса времени шага оптимизации. Внедрение метода KV-компрессии дополнительно сократило время на 14% (совокупное ускорение ×1,45 относительно базового режима). Комплексное применение JIT-компиляции (torch.compile), оптимизированных ядер внимания (xFormers) и вычислений со смешанной точностью (fp16) обеспечило суммарное ускорение в 9,1 раза при снижении пикового потребления видеопамяти на 32% (с 48,2 до 32,6 ГБ); расчетная длительность эпохи сократилась с 24,3 до 2,7 часов. Побочным эффектом перехода на Fp16 стало двукратное сокращение объема кэша признаков в ОЗУ (с 99 до 50 ГБ).

Таблица 3 — Влияние методов оптимизации на эффективность вычислительного процесса

Конфигурация	**Длительность эпохи, мин**	**Пиковое потребление VRAM, ГБ**	**Пропускная способность, изобр/сек**	**Ускорение вычислений**
baseline	1459.4	48.20	0.28	×1.0
+ precompute	1174.4	45.19	0.35	×1.24
+ kv-compression	1009.2	46.18	0.41	×1.45
+ JIT + xformers + fp16	161.2	32.65	2.56	×9.06

В разделе 4.3 проведен сквозной эксперимент по обучению моделей для двух прикладных задач цифровой патологии: сегментации кровеносных сосудов и выявления сосудистой инвазии. Расчет метрик качества генеративной модели (таблица 4) подтверждает высокую близость распределений синтетических и реальных изображений: FID и KID в пространстве Inception-v3 соответствуют уровню современных генеративных моделей цифровой патологии, CLIP-FID близок к нулю, а значение NND синтетических образцов (0,1324) практически совпадает с внутренним расстоянием реальной выборки (0,1317), что доказывает отсутствие прямого копирования обучающих паттернов. Стабильно высокие значения косинусного сходства для трех независимых энкодеров (Virchow-2 (0,8660); UNI-2h (0,8008); H-Optimus-1 (0,7960)) подтверждают сохранение доменно-специфичной семантики заданного визуального условия.

Таблица 4 — Оценка качества сгенерированных изображений. Пояснения к обозначениям: ↓ – меньшее значение соответствует лучшему качеству, ↑ – большее качество, значение KID приведено в масштабе $x10^3$

Аспект качества	**Метрика**	**Пространство признаков**	**Значение метрики**
Согласованность распределений (общевизуальная)	FID ↓	Inception-v3	15.6866
	KID $x10^3$ ↓	Inception-v3	4.9407
Макроструктурная согласованность	CLIP-FID ↓	CLIP ViT-B/32	0.7029
Новизна (отсутствие копирования)	NND (Synth→Real)	Inception-v3	0.1324
	NND (Real→Real)	Inception-v3	0.1317
Качество выборки (реализм и разнообразие)	Precision ↑	Inception-v3	0.8653
	Recall ↑	Inception-v3	0.8819
Сохранение доменно-специфичной семантики	Cosine Similarity ↑	Virchow-2	0.8660
		H-Optimus-1	0.7960
		UNI-2h	0.8008

Стабильно высокие значения метрики косинусного сходства эмбеддингов, наблюдаются сразу для трех независимых фундаментальных моделей-энкодеров (Virchow-2 (0.8660), UNI-2h (0.8008), H-Optimus-1 (0.7960). Это подтверждает, что синтезированные изображения сохраняют доменно-специфичную семантику заданного визуального условия ии располагаются в пространстве признаков целевого домена. Это обеспечивает корректное обогащение обучающей

выборки образцами в стиле дефицитных доменов; наблюдаемый далее прирост метрик прикладных моделей (таблицы 5, 6) косвенно свидетельствует о снижении влияния доменного сдвига при решении downstream-задач.

Прикладная ценность синтетических данных оценивалась по эффекту от расширения обучающих выборок в трех конфигурациях: исходные изображения (source-data); добавление классических геометрических и колориметрических аугментаций (source-data + aug); добавление синтетических изображений, полученных предложенными алгоритмами (source-data + aug + synth). Такая схема изолирует вклад генеративного расширения от эффекта классической аугментации (таблицы 5, 6). Обучение моделей проводилось в режиме кросс-валидации с 10 разбиениями.

Таблица 5 — Качество модели сегментации сосудов (UNet++) при разных конфигурациях обучающей выборки

Конфигурация	**IoU**	**Sensitivity**	**Specificity**
source-data	0.7218 ± 0.0067	0.8778 ± 0.0062	0.8975 ± 0.0056
source-data + aug	0.7292 ± 0.0018	0.8887 ± 0.0043	0.8963 ± 0.0042
source-data + aug + synth	**0.7376 ± 0.0014**	0.8781 ± 0.0036	**0.9094 ± 0.0034**

Таблица 6 — Качество модели классификации участков инвазии (EfficientNet-B1) при разных конфигурациях обучающей выборки

Конфигурация	**F1-score**	**Sensitivity**	**Specificity**
source-data	0.9279 ± 0.0149	0.9477 ± 0.0135	0.958 ± 0.0146
source-data + aug	0.9323 ± 0.0137	0.9616 ± 0.0216	0.9554 ± 0.0064
source-data + aug + synth	**0.9577 ± 0.009**	**0.9721 ± 0.0116**	**0.9744 ± 0.0079**

Результаты экспериментов показывают, что интеграция синтетических изображений в обучающие выборки моделей классификации и сегментации оказывает выраженный регуляризирующий эффект, обеспечивая повышение средних значений метрик качества и снижение их дисперсии. В задаче сегментации кровеносных сосудов среднее значение IoU увеличилось до 0,7376 (±0,0014), а специфичность до 0,9094 (±0,0034), сохранив при этом высокое значение чувствительности алгоритма. В задаче классификации участков инвазии наблюдается комплексный прирост показателей по всем метрикам: рост F1-score до 0,9577 (±0,0090) и показателей чувствительности и специфичности до 0,9721 и 0,9744 соответственно.

В **заключении** перечислены основные результаты работы:

1. Разработан метод адаптивного синтеза гистологических изображений высокого разрешения, основанный на гибридном объединении архитектуры латентных диффузионных трансформеров (DiT), механизмов визуального обусловливания и алгоритмов сжатия матриц внимания (KV-компрессии). Метод формирует вычислительно эффективный и модульный архитектурный фундамент, обеспечивающий строгий контроль над морфологической достоверностью генерируемых структур и позволяющий независимо применять компоненты генеративного конвейера при обработке гигапиксельных изображений.
2. Разработан метод доменной специализации латентных диффузионных моделей, отличающийся двухэтапной декомпозицией процесса: применением прогрессивного обучения с фиксацией оптического увеличения для адаптации морфологических представлений и библиотеки низкоранговых адаптеров (LoRA) для параметрической коррекции доменного стиля. Метод позволяет масштабируемо настраивать генеративные модели под целевые наборы данных без полного переобучения базовой архитектуры и при сохранении визуально обусловленной геометрии структур, эффективно компенсируя ковариантный сдвиг.
3. Разработан алгоритм локального контекстно-ориентированного встраивания объектов, отличающийся пространственно-детерминированной бесшовной интеграцией целевого объекта в контекст фонового изображения на уровне латентных представлений с процедурой автоматического колориметрического согласования. Алгоритм обеспечивает естественное объединение структур без привлечения дополнительных обучаемых архитектурных модулей и сложной пространственной разметки.
4. Разработаны программно-алгоритмический комплекс адаптивного синтеза с модульной архитектурой и гибридная методология многофакторной оценки качества, всесторонне характеризующая статистическую согласованность, семантическую достоверность и генеративное разнообразие выборок. Экспериментально доказано, что применение разработанных решений для обогащения обучающих выборок обеспечивает повышение метрик качества прикладных нейросетевых моделей сегментации и классификации в условиях классового дисбаланса.

Публикации автора по теме диссертации

1. Timakova A., **Ananev V.**, Fayzullin A., Makarov V., Ivanova E., Shekhter A., Timashev P. Artificial intelligence assists in the detection of blood vessels in whole slide images: practical benefits for oncological pathology // Biomolecules. – 2023. – Vol. 13, No. 9. – P. 1327. DOI: https://doi.org/10.3390/biom13091327.
2. Timakova A., **Ananev V.**, Fayzullin A., Zemnuhov E., Rumyantsev E., Zharov A., Zharkov N., Zotova V., Shchelokova E., Demura T., Timashev P., Makarov V. LVI-PathNet: Segmentation-classification pipeline for detection of lymphovascular invasion in whole slide images of lung adenocarcinoma // Journal of Pathology Informatics. – 2024. – Vol. 15. – P. 100395. DOI: https://doi.org/10.1016/j.jpi.2024.100395.
3. Timakova A., Fayzullin A., **Ananev V.**, Zemnuhov E., Alfimov V., Baranov A., Smirnova Y., Shatalov V., Konukhova N., Karpulevich E., Timashev P., Makarov V. Computer Vision-Assisted Spatial Analysis of Mitoses and Vasculature in Lung Cancer // Journal of Clinical Medicine. – 2025. – Vol. 14, No. 21. – P. 7526. DOI: https://doi.org/10.3390/jcm14217526.
4. **Ананьев В. В.**, Ефимов Э. В., Земнухов Е. С., Тимакова А. А., Карпулевич Е. А. Метод адаптивного синтеза изображений высокого разрешения на основе диффузионных трансформеров // Труды Института системного программирования РАН. – 2026. – Т. 38, вып. 4.
5. Земнухов Е. С., **Ананьев В. В**., Макаров В. А. “Программный модуль подготовки гистологических изображений для задач машинного обучения” / Свидетельство о государственной регистрации программы для ЭВМ, рег. №2024663138 от 04.06.2024 – Российская Федерация, 2024.

Ананьев Владислав Валерьевич

Метод и программные средства адаптивного синтеза изображений высокого разрешения на основе диффузионных моделей при работе с гигапиксельными изображениями

Автореф. дис. на соискание ученой степени канд. технических наук

Подписано в печать __.__.____. Заказ № _____
Формат 60×90 / 16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____