DOI: 10.15514/ISPRAS-2029-38(4)-12

Метод адаптивного синтеза изображений высокого разрешения на основе диффузионных трансформеров

[1] *В.В. Ананьев, ORCID: 0000-0002-5070-8117 <novisp53@ispras.ru>*
[1] *Э.В. Ефимов, ORCID: 0009-0009-1058-5018 <e.efimov@ispras.ru>*
[1] *Е.С. Земнухов, ORCID: 0009-0003-0067-0063 <e.zemnukhov@ispras.ru>*
[2] *А.А. Тимакова, ORCID: 0009-0004-5807-4501 <atimakova96@gmail.com>*
[1] *Е.А. Карпулевич, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

[2] *Сеченовский Университет, Россия, 119048, город Москва, Трубецкая ул, д. 8, стр. 2.*

Аннотация. В работе рассматривается задача адаптивного синтеза изображений высокого разрешения для цифровой патологии. Предложен генеративный конвейер на базе латентного диффузионного трансформера, объединяющий кодирование изображений в латентное пространство, визуальное обусловливание признаками фундаментальной модели UNI2-h, архитектурную оптимизацию слоев внимания за счет KV-компрессии и доменную специализацию с использованием низкоранговых адаптеров LoRA. Описан алгоритм контекстно-ориентированного встраивания объектов, который совмещает пространственно-хроматическое выравнивание, семантически обусловленную диффузию из промежуточного латентного состояния и морфологическое смешивание границ. Программная реализация построена как модульный комплекс из контуров ввода-вывода, латентно-признакового кодирования, генерации и подготовки данных для прикладных моделей. Экспериментальная верификация выполнена на задачах сегментации сосудистых структур и классификации участков лимфоваскулярной инвазии; результаты показывают, что направленное добавление синтетических изображений может повышать качество прикладных моделей в условиях дефицита данных. Эксперименты выполнены с использованием базы из 449 полнослайдовых изображений (WSI), состоящей из выборок DHMC, LCNOV и NLST. Синтетическое расширение выборки включает 5172 изображений для сегментации и 4520 изображений для классификации. Используемые аппаратно-программные оптимизации позволили сократить длительность эпохи обучения и пиковое потребление VRAM, добавление синтетических данных повысило значение метрики IoU модели UNet++ и значение метрики F1-score модели EfficientNet-b1.

Ключевые слова: латентная диффузионная модель; диффузионный трансформер; синтез изображений; визуальное обусловливание; адаптер LoRA; технология inpainting; гигапиксельные изображения; цифровая патология.

Для цитирования: 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 ИСП РАН, 2026, том 38, вып. 4, часть 1, стр. 225–238. DOI: 10.15514/ISPRAS-2026-38(4)-12.

Благодарности: Результаты получены с использованием услуг Центра коллективного пользования Института системного программирования им.В.П.Иванникова РАН – ЦКП ИСП РАН.

Adaptive high-resolution image synthesis based on diffusion transformers

[1] *V. Ananev, ORCID: 0000-0002-5070-8117 <novisp53@ispras.ru>*
[1] *E. Efimov, ORCID: 0009-0009-1058-5018 <e.efimov@ispras.ru>*
[1] *E. Zemnukhov, ORCID: 0009-0003-0067-0063 <e.zemnukhov@ispras.ru>*
[2] *A. Timakova, ORCID: ORCID: 0009-0004-5807-4501 <atimakova96@gmail.com>*
[1] *E. Karpulevich, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

[2] *Sechenov University, Trubetskaya st., 8 building 2., Moscow, 119048, Russia.*

Abstract. The paper addresses adaptive high-resolution image synthesis for digital pathology. The problem is caused by the shortage of annotated histology data, expensive expert labeling, covariate shift between whole-slide imaging devices and severe class imbalance in tasks involving minor morphological structures. We propose a generative pipeline based on a latent diffusion transformer. The pipeline combines image encoding into a latent space, visual conditioning using features extracted by the UNI2-h pathology foundation encoder, attention-layer optimization through key-value compression and domain specialization with low-rank LoRA adapters. We also present a context-oriented object insertion algorithm that performs spatial and chromatic alignment, semantically conditioned diffusion from an intermediate latent state and boundary-aware morphological blending. The software implementation is organized as a modular system with independent input-output, latent-feature encoding, generative and downstream data preparation components. Experiments are designed for blood vessel segmentation and lymphovascular invasion classification on whole-slide image fragments. The results confirm that targeted synthetic data generation can improve downstream model quality under data scarcity and class imbalance. The experiments were carried out using a database of 449 WSIs consisting of the DHMC, LCNO and NLST datasets, where 216 WSIs are annotated; vessel segmentation used 8212 masks converted into 5764 images, while invasion classification used 1875 images. The synthetic extension included 5172 images for segmentation and 4520 images for classification. The applied hardware and software optimizations reduced the epoch time from 1459.4 to 161.2 min and peak VRAM consumption from 48.20 to 32.65 GB; FID was 15.6866, KID×10³ was 4.9407, and the addition of synthetic data increased the UNet++ IoU to 0.7376±0.0014 and the EfficientNet-b1 F1-score to 0.9577±0.0090.

Keywords: latent diffusion model; diffusion transformer; image synthesis; visual conditioning; LoRA adapter; inpainting; gigapixel images; digital pathology.

For citation: Ananev V., Efimov E., Zemnukhov E., Timakova A., Karpulevich E. Adaptive high-resolution image synthesis based on diffusion transformers. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 4, part 1, pp. 225-238 (in Russian). DOI: 10.15514/ISPRAS-2026-38(4)-12.

Acknowledgements. The results were obtained using the services of the Ivannikov Institute for System Programming (ISP RAS) Data Center.

1. Введение

Глубокие нейронные сети (ГНС) стали базовым инструментом решения задач компьютерного зрения, включая классификацию, сегментацию и поиск объектов на медицинских изображениях. Однако их практическое применение в цифровой патологии сталкивается с рядом системных ограничений. Во-первых, обработка гистологических изображений сверхвысокого разрешения требует значительных вычислительных ресурсов. Во-вторых, качество моделей существенно зависит от объема и разнообразия аннотированных данных. В-третьих, экспертная разметка структур на изображениях трудоемка, плохо масштабируется и требует участия специалистов предметной области.

Эти ограничения особенно заметны при работе с цифровыми полнослайдовыми изображениями (whole-slide images, WSI). Одно изображение WSI может иметь размер

порядка $10^5 \times 10^5$ пикселей и хранится как многомасштабная иерархическая пирамида. Для обучения ГНС такое изображение обычно декомпозируется на фрагменты (тайлы) фиксированного разрешения, соответствующие выбранному уровню оптического увеличения. При этом на уровне данных возникают дополнительные источники нестабильности: различия между сканерами, источниками освещения, матрицами камер, алгоритмами автофокуса, балансом белого и протоколами окрашивания. В результате распределение признаков в обучающей выборке может заметно отличаться от распределения в целевой среде.

Отдельную проблему составляет классовый дисбаланс. В рассматриваемой задаче анализа сосудистых структур на 216 WSI было размечено 8212 объектов сосудистого русла, при этом только 216 объектов, то есть менее 3%, относились к целевому минорному классу. Классические аугментации, такие как повороты, отражения, масштабирование, цветовая нормализация и случайные колориметрические преобразования, повышают устойчивость модели, но не создают принципиально новых топологических конфигураций объектов. Следовательно, они не полностью компенсируют семантический дефицит обучающей выборки.

Перспективным направлением является использование генеративных моделей, способных формировать новые изображения с сохранением морфологической достоверности и заданного доменного стиля. В настоящей статье описывается метод адаптивного синтеза изображений высокого разрешения на основе латентных диффузионных трансформеров. В отличие от моделей с текстовым управлением (text-to-image), предложенный подход ориентирован на визуальное обусловливание: управление генерацией осуществляется не текстовым описанием, а признаками референсного изображения, извлекаемыми специализированным энкодером цифровой патологии.

В работе предложен и экспериментально проверен подход к синтезу гистологических изображений высокого разрешения на основе латентной диффузионной модели с трансформерной архитектурой. Особое внимание уделено вычислительной эффективности такого конвейера: для снижения затрат на обработку изображений большого разрешения используется сжатие матриц ключей и значений в слоях внимания. Для адаптации модели к особенностям отдельных наборов данных применяется набор низкоранговых LoRA-адаптеров (Low-Rank Adaptation), позволяющий учитывать различия в стиле изображений без полного переобучения базовой модели. Отдельной частью работы является алгоритм контекстного встраивания объектов, предназначенный для направленного расширения выборок с редкими классами. Разработанные методы объединены в модульный программно-алгоритмический комплекс; его применение оценивалось по влиянию синтетических данных на качество прикладных моделей сегментации и классификации.

2. Обзор существующих решений

Ранние подходы к синтезу изображений в компьютерном зрении во многом опирались на генеративно-состязательные сети (generative adversarial networks, GAN) [1] и вариационные автокодировщики (variational autoencoders, VAE) [2]. GAN-модели способны формировать визуально реалистичные изображения, но их обучение часто нестабильно, а результат чувствителен к балансу между генератором и дискриминатором. Кроме того, в практических задачах они могут демонстрировать режимный коллапс, то есть генерировать ограниченное число типовых вариантов вместо покрытия всего распределения данных. VAE, напротив, имеют более устойчивую вероятностную постановку, но часто дают сглаженные изображения и хуже сохраняют мелкие текстурные детали, важные для гистологии.

Вероятностные диффузионные модели (denoising diffusion probabilistic models, DDPM) [3] изменили практику генеративного моделирования. Их обучение сводится к восстановлению данных из последовательности зашумленных состояний, а генерация выполняется как

постепенное обращение процесса зашумления. Такой подход обеспечивает более стабильную оптимизацию и хорошо масштабируется на сложные распределения изображений. Однако прямое применение диффузии в пиксельном пространстве вычислительно затратно, особенно для изображений высокого разрешения.

Латентные диффузионные модели (latent diffusion models, LDM) [4] уменьшают вычислительную нагрузку за счет переноса диффузионного процесса из пространства пикселей в сжатое латентное пространство автокодировщика. В этой постановке модель работает не с полным изображением, а с компактным представлением, из которого затем декодируется итоговый результат. Следующим шагом стало использование диффузионных трансформеров (diffusion transformers, DiT) [5], где сверточная U-Net-магистраль заменяется трансформером, работающим с латентными патчами. Такая архитектура лучше моделирует дальние зависимости между областями изображения, что важно для сложноструктурированных данных.

Для масштабирования диффузионных трансформеров к высоким разрешениям применяются архитектурные оптимизации. В работе [6] показано, что сжатие матриц ключей и значений в слоях внимания модели PixArt-Sigma снижает стоимость вычислений и позволяет работать с изображениями сверхвысокого разрешения. Для цифровой патологии особенно интересны специализированные модели, например PixCell [7], где диффузионный трансформер обучается на большом корпусе H&E-окрашенных WSI-тайлов и использует визуальное обусловливание вместо текстового описания.

Выбор механизма обусловливания является принципиальным. В задачах общего синтеза изображений широко применяется текстовое управление, однако для гистологических структур естественный язык недостаточно точно задает локальную геометрию, тип окраски, структуру ткани и взаимное расположение объектов. В связи с этим более надежным решением становится управление генерацией через изображения (image-to-image), при котором используются векторы признаков, извлекаемые специализированными фундаментальными моделями цифровой патологии. В качестве таких моделей могут использоваться UNI [8], UNI2-h и Virchow-2 [9], обученные на больших коллекциях гистологических данных и формирующие информативные эмбеддинги тканевых структур.

Доменная адаптация генеративных моделей может осуществляться путем полного дообучения, однако этот подход вычислительно затратен и на малых выборках часто приводит к переобучению, а также к потере ранее усвоенных признаков. Альтернативой является низкоранговая адаптация LoRA [10], при которой базовые веса модели фиксируются, а обучаемое изменение задается произведением двух матриц малого ранга. Для рассматриваемой задачи это позволяет хранить отдельный адаптер для каждого домена или набора данных, не создавая отдельную копию всей модели.

Существующие исследования позволяют получить набор из ряда важных компонентов для работы с гистологическими изображениями. При этом для практического применения в цифровой патологии требуется единый конвейер, объединяющий эти компоненты, поддерживающий высокие разрешения, работу с WSI-тайлами, доменную специализацию и автоматизированное формирование аннотированных наборов данных для прикладных моделей.

3. Метод адаптивного латентного синтеза

3.1 Обобщенная схема генеративного конвейера

Предложенный метод строится как латентный генеративный конвейер с тремя основными компонентами: подсистемой кодирования и декодирования изображений, диффузионным трансформером и блоком визуального обусловливания. Входное изображение или референсный фрагмент WSI переводится в латентное пространство с помощью

автокодировщика SD3-VAE. Диффузионный трансформер выполняет обратный процесс расшумления в этом пространстве, а управляющее условие формируется энкодером UNI2-h. Схема конвейера приведена на рис. 1.

В отличие от классических сверточных генеративных моделей, DiT работает с последовательностью латентных патчей. Это позволяет использовать механизм внимания для учета глобального контекста и повышает устойчивость синтеза при формировании сложных тканевых паттернов. Перенос диффузии в латентное пространство уменьшает размерность вычислений, что критично при переходе от тайлов 256×256 к разрешениям 512×512 и 1024×1024 пикселей.

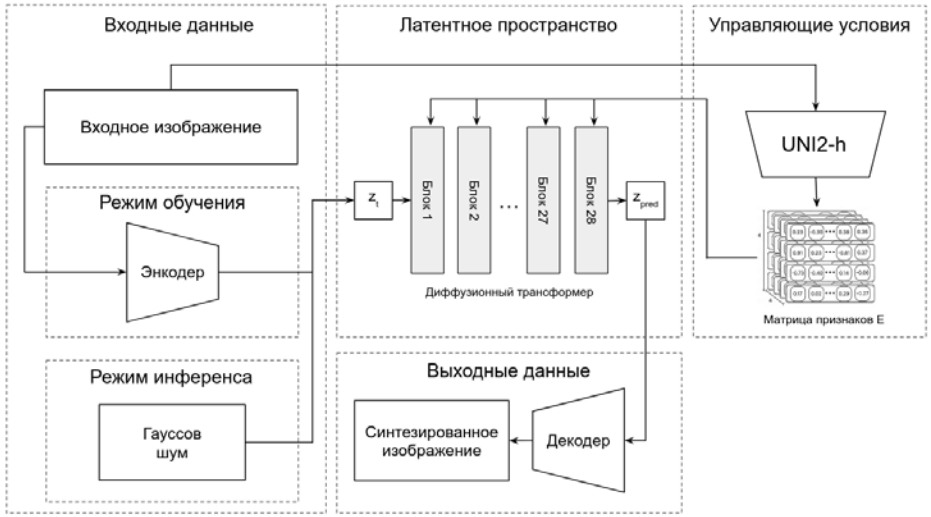

Рис. 1. Обобщенная схема латентного генеративного конвейера.
Fig. 1. General scheme of the latent generative pipeline.

3.2 Диффузионный процесс и визуальное обусловливание

Пусть x обозначает исходный тайл WSI, а E_VAE и D_VAE – кодировщик и декодировщик автокодировщика. Латентное представление задается как:

$$z_0 = E\_VAE(x), \quad x\_hat = D\_VAE(z_0)$$

Прямой процесс диффузии последовательно добавляет гауссов шум к латентному представлению. Для произвольного шага t он может быть записан в стандартной форме:

$$q(z_t \mid z_0) = N(sqrt(alpha\_bar\_t)\ z_0,\ (1 - alpha\_bar\_t)\ I)$$

Здесь z_t – зашумленное латентное представление на шаге t, $alpha\_bar\_t$ – накопленный коэффициент расписания шума, I – единичная матрица. Обратный процесс аппроксимируется моделью p_theta, которая восстанавливает менее зашумленное состояние с учетом управляющего признакового вектора e:

$$p\_theta(z_t - 1 \mid z_t, e), \qquad e = E\_UNI(x\_ref)\ in\ R^{16\times1536}$$

Вектор или матрица признаков e извлекается из референсного изображения x_ref с помощью UNI2-h и передается в трансформерные блоки через механизм перекрестного внимания. В результате генерация управляется не текстовой инструкцией, а признаковым описанием тканевой структуры. Это снижает неоднозначность управления и делает метод более

подходящим для медицинских изображений, где важны локальная морфология, цветовые характеристики и пространственный контекст.

3.3 Метод архитектурной оптимизации

Основным вычислительным узким местом трансформеров DiT является механизм внимания. При N латентных токенах стандартная сложность вычисления матрицы внимания имеет порядок $O(N^2)$. Для изображений разрешением 1024×1024 и выше это приводит к резкому росту потребления видеопамяти. Для снижения нагрузки в выбранных слоях перекрестного внимания (cross-attention) применяется сжатие кодирующих контекст матриц “ключей” (Key, K) и “значений” (Value, V), при сохранении полного набора “запросов” (Query, Q), отвечающих за текущее генерируемое представление. В этой схеме число токенов в матрицах K и V уменьшается в sr^2 раз, где sr - коэффициент пространственной редукции:

$$Attention(Q, K, V) = softmax(Q\ K\text{^}T / sqrt(d))\ V, \qquad K,V \text{ -> } Compress\_sr(K,V).$$

$$O(N^2) \text{ -> } O(N * N / sr^2).$$

На практике компрессия применяется не во всех блоках, а в наиболее ресурсоемких средних и поздних слоях трансформера, например в блоках 14-28. Это позволяет сохранить способность модели учитывать глобальные признаки через Query, но уменьшить размерность операций с ключами и значениями. Такой выбор является компромиссом между качеством синтеза и ресурсной эффективностью.

3.4 Стратегия прогрессивного обучения

Для стабилизации обучения используется прогрессивное наращивание разрешения. Сначала модель обучается на тайлах 256×256, затем переносится на 512×512 и далее на 1024×1024 пикселей. Важным условием является сохранение фиксированного оптического увеличения: переход между разрешениями не должен смешивать морфологические признаки, соответствующие разным физическим масштабам. Результатом является семейство базовых моделей *W0(256)*, *W0(512)* и *W0(1024)*, которые работают с изображениями различного разрешения, но сохраняют согласованность в пространстве признаков.

Прогрессивное обучение уменьшает вероятность расходимости функции потерь на высоких разрешениях и позволяет использовать веса предыдущего этапа как хорошую начальную точку. Для цифровой патологии это особенно важно, так как структура ткани имеет одновременно локальные микропаттерны и более крупные пространственные зависимости.

3.5 Метод доменной специализации моделей

Ковариантный сдвиг между наборами данных в цифровой патологии часто обусловлен не только биологическими различиями, но и аппаратными профилями сканеров, цветопередачей, качеством фокусировки, процедурой окрашивания и подготовкой препарата. Полное дообучение генеративной модели под каждый новый домен вычислительно неэффективно. Поэтому предлагается использовать библиотеку LoRA-адаптеров, где один адаптер соответствует одному домену, например отдельному набору данных или лабораторному протоколу.

$$W\_domain = W0 + DeltaW, \qquad DeltaW = B\ A, \qquad rank(DeltaW) = r << min(d\_in, d\_out)$$

Базовая модель $W0$ остается замороженной, а обучаются только малые матрицы A и B. Во время инференса адаптер может подключаться динамически. Такой подход позволяет переносить доменный стиль без переобучения полного набора параметров и снижает риск катастрофического забывания. В рамках предлагаемого метода LoRA настраивается прежде всего в слоях перекрестного внимания, так как именно они связывают визуальное условие с процессом генерации.

3.6 Алгоритм контекстно-ориентированного встраивания объектов

Для компенсации дисбаланса минорных классов разработан алгоритм контекстно-ориентированного встраивания объектов. Его задача - не просто вставить фрагмент с целевой структурой в фон, а согласовать объект с окружающей тканью так, чтобы итоговое изображение могло использоваться для обучения прикладной модели. Алгоритм состоит из трех этапов: пространственно-хроматического выравнивания, семантически обусловленной диффузии из промежуточного латентного состояния и морфологической коррекции границ (рис. 2).

Рис. 2. Схема контекстно-ориентированного встраивания объектов.
Fig. 2. Context-oriented object insertion scheme.

На первом этапе выбирается фоновое изображение T и донорский объект D с маской M. Выполняется подбор координат, случайная геометрическая трансформация, а также цветовая гармонизация за счет выравнивания статистических моментов в перцептуально однородной цветовой модели CIELAB, которая позволяет независимо обрабатывать каналы яркости и цвета. На основе полученных трансформированных версий донора D' и маски M' формируется предварительный композит:

$$X_{comp} = (1 - M') * T + M' * D'$$

На втором этапе композит кодируется в латентное пространство и зашумляется до промежуточного шага t. Обратная диффузия запускается не из чистого шума, а из z_t, что ограничивает глубину преобразования и сохраняет общую геометрию сцены. Следование семантическому условию регулируется коэффициентом CFG, а по завершении процесса обратной диффузии и декодирования формируется сгенерированное изображение X_{gen}.

На третьем этапе вычисляются поля расстояний до границы маски, на основе которых формируется карта воздействия A. Ее значения максимальны на границе и экспоненциально затухают вглубь как донора, так и фона. Итоговое изображение вычисляется как взвешенная комбинация композита и результата диффузии:

$$X_{final} = (1 - A) * X_{comp} + A * X_{gen}$$

Такое смешивание концентрирует генеративное воздействие в узкой зоне вдоль шва и уменьшает риск артефактов внутри донорского объекта. Полную математическую модель метода можно свернуть в единую функцию граничного смешивания $B(X_{gen}, X_{comp}, M', e, t, s)$, где s – коэффициент CFG, а e – визуальное условие (семантические признаки), извлекаемое

фундаментальным энкодером. При этом X_{gen} формируется на основе латентного представления композита и указанных признаков e.

4. Программная реализация

Программный комплекс реализован как набор слабосвязанных функциональных контуров. Такая архитектура позволяет независимо изменять компоненты чтения WSI, кодирования признаков, генерации и формирования обучающих наборов данных. Общая программная схема приведена на рис. 3.

Рис. 3. Модульная программная архитектура комплекса.
Fig. 3. Modular software architecture of the system.

Контур ввода-вывода отвечает за потоковое чтение WSI и асинхронную декомпозицию слайдов на тайлы фиксированного разрешения. Это позволяет работать с изображениями, которые не помещаются в оперативную память или видеопамять целиком. Контур латентно-признакового кодирования объединяет автокодировщик SD3-VAE с моделью UNI2-h: первый формирует латентное представление изображения, второй – семантическое визуальное условие. Оба типа представлений могут вычисляться заранее и сохраняться в кэше.

Генеративный контур построен на базе модифицированной абстракции DiffusionPipeline библиотеки diffusers. В него входят диффузионный трансформер, библиотека LoRA-адаптеров, механизм выбора домена и процедуры генерации в режимах вариативного синтеза и контекстного встраивания. Для распределенного обучения используется DistributedDataParallel. Контур прикладных задач обеспечивает сопряжение с надстройкой huggeron и автоматическое формирование структурированных наборов данных, включающих изображения, маски, метки классов и служебные метаданные.

Для повышения производительности реализованы несколько уровней аппаратно-программной оптимизации. Предварительное вычисление латентных представлений и UNI2-h эмбеддингов снижает число повторных обращений к дисковой подсистеме. Многопоточное RAM-кэширование уменьшает время простоя GPU при обучении. В свою очередь, JIT-компиляция, применение алгоритмов памяти-эффективного внимания (библиотека xFormers) и использование форматов чисел с плавающей запятой половинной точности (fp16/bf16) приводят к снижению вычислительной стоимости отдельных операций и позволяют увеличить размер мини-партии.

Фиксированный формат межконтурных интерфейсов обеспечивает асинхронность конвейера и упрощает замену отдельных модулей.

5. Эксперименты и результаты

5.1 Методология экспериментов

Экспериментальная проверка выполнялась на аппаратном комплексе с одним графическим ускорителем NVIDIA A100. Программная среда включала CUDA, PyTorch, diffusers, xformers и вспомогательные модули для подготовки WSI-тайлов.

Тестирование проводилось на трех выборках WSI: DHMC, LCNOV и NLST. Разделение на обучающую и контрольную части выполнялось на уровне слайдов, чтобы исключить утечку визуально близких тайлов между выборками. Оценка прикладных моделей выполнялась по схеме 10-кратной кросс-валидации. Состав исходных клинических выборок приведен в табл. 1.

Табл. 1. Состав исходных клинических выборок.
Table 1. Source clinical datasets.

Выборка	Аннотированные WSI	Всего WSI
DHMC	105	143
LCNOV	54	161
NLST	57	145
Всего	**216**	**449**

Для задачи сегментации использовано 8212 масок кровеносных сосудов, размеченных на 216 WSI и преобразованных в 5764 изображений. Для задачи классификации лимфоваскулярной инвазии отобрано 1875 изображений двух классов: «инвазия» и «чистый сосуд». Синтетические изображения формировались двумя способами: вариативным синтезом на основе референсов и контекстно-ориентированным встраивания. Объемы исходных и синтетических данных для прикладных задач приведены в табл. 2.

Табл. 2. Объем данных по прикладным задачам.
Table 2. Data volume for downstream tasks.

Прикладная задача	Исходные изображения	Синтетические изображения
Сегментация сосудов	5764	5172
Классификация инвазии	1875	4520

Качество генерации оценивалось гибридной системой метрик. Визуальная близость распределений реальных и синтетических данных оценивалась через метрики FID [15] и KID [16] в пространстве Inception-v3. Для учета доменной специфики дополнительно использовались метрики Frechet distance и Kernel distance в пространствах Virchow-2 и UNI2-h. Разнообразие и покрытие распределения контролировались метриками Precision/Recall [17]. Риск прямого копирования обучающих образцов оценивался через расстояние до ближайшего реального образца (NND), а соответствие визуальному условию – через косинусное сходство эмбеддингов.

Практическая ценность синтетических данных оценивалась через качество прикладных моделей: UNet++ для сегментации сосудов [11] и EfficientNet-b1 для классификации участков инвазии [12]. Для сегментации использовались метрики IoU, Sensitivity, Specificity и Recall, для классификации – Precision, Recall, F1-score, Sensitivity и Specificity.

5.2 Оценка вычислительной эффективности оптимизаций

Сравнительный анализ базового режима генерации векторов признаков «на лету» (baseline) и предложенного подхода с предварительным вычислением латентных векторов SD3-VAE и эмбеддингов UNI2-h (optimized) показал сокращение времени обучения на одну эпоху на 19,5% при росте пропускной способности конвейера данных в 1,24 раза и кратном снижении

разброса времени шага оптимизации, что обусловлено устранением зависимости конвейера от дисковой подсистемы.

Внедрение метода kv-компрессии дополнительно сокращает время режима optimized на 14% (совокупное ускорение ×1,45 по отношению к базовому режиму). Комплексное применение JIT-компиляции (torch.compile), оптимизированных ядер внимания (xformers) и вычислений со смешанной точностью (fp16) обеспечило суммарное ускорение расчетов в 9,1 раза при снижении пикового потребления видеопамяти на 32% (с 48,20 до 32,65 ГБ). Расчетная длительность эпохи сократилась с 24,3 до 2,7 часа, что позволяет увеличить размер мини-партии (batch size) при обучении модели. Побочным эффектом перехода на fp16 стало двукратное сокращение объема кэша предвычисленных признаков в оперативной памяти (с 99 до 50 ГБ). Количественные результаты приведены в табл. 3.

Табл. 3. Влияние методов оптимизации на эффективность вычислительного процесса.
Table 3. Effect of optimization methods on computational efficiency.

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 + zformers + fp16	161.2	32.65	2.56	×9.06

5.3 Качество синтезированных изображений

Оценка качества генерации выполнялась на уровне распределений признаков и на уровне локального сходства с управляющим условием. Целевым результатом является не минимизация одной метрики, а согласованное выполнение нескольких требований: синтетические изображения должны быть близки к реальным по распределению признаков, достаточно разнообразны, не копировать обучающие образцы и сохранять заданный морфологический контекст. Сводные значения метрик качества синтезированных изображений приведены в табл. 4.

5.4 Влияние синтетических данных на прикладные модели

Прикладная оценка выполнялась в трех конфигурациях обучающей выборки: только исходные реальные изображения (src-data), реальные изображения с классическими аугментациями (src-data + aug), реальные изображения с классическими аугментациями и синтетическими данными (src-data + aug + gen). Такая схема позволяет отделить эффект генеративного расширения данных от эффекта стандартной аугментации.

Для сегментации сосудов использовалась модель UNet++, для классификации участков лимфоваскулярной инвазии – EfficientNet-b1. Наиболее важным практическим результатом является повышение чувствительности к минорному классу при сохранении приемлемой специфичности. Значения метрик для обеих прикладных задач приведены в табл. 5 и 6.

Результаты экспериментов показывают, что интеграция синтетических изображений в обучающие выборки моделей классификации и сегментации оказывает выраженный регуляризирующий эффект, обеспечивая статистически значимое повышение качества и снижение дисперсии ошибок. 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. Генеративный модуль выступает не как самостоятельная демонстрационная система, а как компонент конвейера подготовки обучающих выборок. Его эффективность должна оцениваться через прикладные (downstream) задачи, поскольку именно они определяют практическую ценность сгенерированных данных для конечного программного комплекса цифровой патологии.

Табл. 4. Оценка качества сгенерированных изображений.
Table 4. Quality assessment of generated images.

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

Табл. 5. Качество модели сегментации сосудов UNet++ при разных конфигурациях обучающей выборки.
Table 5. UNet++ vessel segmentation quality for different training sets.

Конфигурация	IoU	Sensitivity	Specificity	Комментарий
src-data	0.7218±0.0067	0.8778±0.0062	0.8975±0.0056	базовое обучение
src-data + aug	0.7292±0.0018	0.8887 ± 0.0043	0.8963±0.0042	классические аугментации
src-data + aug + gen	**0.7376±0.0014**	0.8781 ± 0.0036	**0.9094±0.0034**	добавлены синтетические изображения

Табл. 6. Качество модели классификации участков инвазии EfficientNet-b1.
Table 6. EfficientNet-b1 invasion classification quality.

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

6. Обсуждение

Предложенный подход применим в задачах, где изображения имеют высокое разрешение, сложную внутреннюю структуру и выраженный дефицит разметки. В цифровой патологии это задачи сегментации сосудов, поиска митозов, выявления редких морфологических паттернов, анализа инвазии и подготовки обучающих выборок для многоэтапных диагностических конвейеров. Потенциально та же логика может быть перенесена на другие классы изображений сверхвысокого разрешения, например рентгенографию, КТ-срезы с высоким разрешением, микроскопию материалов и спутниковые снимки. При этом для каждого нового домена потребуется собственный механизм проверки реалистичности и прикладной полезности синтетических данных.

Ключевое преимущество метода заключается в сочетании визуального обусловливания и модульной доменной адаптации. Визуальное условие позволяет работать без трудоемких текстовых описаний морфологических структур, а LoRA-адаптеры дают возможность быстро менять доменный стиль. Это особенно важно при работе с данными разных лабораторий, где вариативность изображений часто связана не с биологией, а с техническими характеристиками процесса оцифровки.

Основные ограничения связаны с необходимостью строгой валидации синтетики. Метрики генерации, такие как FID и KID, полезны, но сами по себе не гарантируют клиническую или морфологическую корректность. Поэтому в методологию включены доменные эмбеддинги Virchow-2 и UNI2-h, NND для контроля копирования и прикладные метрики моделей сегментации и классификации. Дополнительно требуется экспертный визуальный контроль части синтетических изображений, особенно если они используются для редких классов или клинически значимых структур.

Второе ограничение связано с точностью контекстного встраивания. Даже при цветовой гармонизации в пространстве CIELAB и морфологического смешивания границ возможно появление тонких артефактов, которые не всегда заметны глобальным метрикам. Поэтому дальнейшее развитие алгоритма должно включать автоматическое выявление артефактов, более гибкий подбор фонового контекста и сравнение разных стратегий выбора промежуточного шага диффузии t.

Наконец, программная реализация требует контроля воспроизводимости. Для практического использования необходимо фиксировать версии моделей, параметры адаптеров, конфигурации генерации, случайные зерна, версии библиотек и сведения о происхождении исходных изображений. Это позволит использовать синтетические данные в воспроизводимых научных экспериментах и упростит аудит результатов.

7. Заключение

В статье представлен метод адаптивного синтеза изображений высокого разрешения на основе латентных диффузионных трансформеров для задач цифровой патологии. Метод объединяет латентную диффузию, визуальное обусловливание признаками специализированного энкодера, KV-компрессию слоев перекрестного внимания, прогрессивное обучение и LoRA-адаптацию доменного стиля.

Разработан алгоритм контекстно-ориентированного встраивания объектов, предназначенный для направленного расширения минорных классов. Алгоритм сочетает пространственно-хроматическое выравнивание, диффузионную генерацию из промежуточного латентного состояния и морфологическую коррекцию границ. Такой подход позволяет формировать изображения, пригодные для обучения прикладных моделей, а не только для визуальной демонстрации генеративной модели.

Описана модульная программная реализация, включающая контуры ввода-вывода, латентно-признакового кодирования, генерации и подготовки данных для прикладных задач. Экспериментальная проверка на выборках DHMC, LCNOV и NLST показала, что

предложенные оптимизации обеспечивают ускорение вычислений до ×9,06 и снижение пикового потребления VRAM до 32,65 ГБ, а использование синтетических данных повышает качество моделей сегментации сосудистых структур и классификации лимфоваскулярной инвазии в условиях дефицита разметки и выраженного классового дисбаланса.

Список литературы / References

[1]. Goodfellow I. et al. Generative adversarial networks. Communications of the ACM, 2020, vol. 63, no. 11, pp. 139-144.
[2]. Kingma D. P., Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.
[3]. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models, Advances in neural information processing systems, 2020, vol. 33, pp. 6840-6851.
[4]. Rombach R. et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684-10695.
[5]. Peebles W., Xie S. Scalable diffusion models with transformers. Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4195-4205.
[6]. Chen J. et al. Weak-to-strong training of diffusion transformer for 4k text-to-image generation. European Conference on Computer Vision, pp. 74-91.
[7]. Yellapragada S. et al. PixCell: A generative foundation model for digital histopathology images. arXiv: 2506.05127 v1, 2025.
[8]. Chen R. J. et al. Towards a general-purpose foundation model for computational pathology. Nature medicine, 2024, vol. 30, no. 3, pp. 850-862.
[9]. Vorontsov E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine, 2024, vol. 30, no. 10, pp. 2924-2935.
[10]. Hu E. J. et al. Lora: Low-rank adaptation of large language models. Iclr., 2022, vol. 1, no. 2, p. 3.
[11]. Zhou Z. et al. Unet++: A nested u-net architecture for medical image segmentation. International workshop on deep learning in medical image analysis. Cham: Springer International Publishing, 2018, pp. 3-11.
[12]. Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks //International conference on machine learning. PMLR, 2019, pp. 6105-6114.
[13]. Heusel M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 2017, vol. 30.
[14]. Bińkowski M. et al. Demystifying mmd gans. arXiv preprint arXiv:1801.01401, 2018.
[15]. Kynkäänniemi T. et al. Improved precision and recall metric for assessing generative models. Advances in neural information processing systems, 2019, vol. 32.
[16]. Timakova A. et al. 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.
[17]. Timakova A. et al. 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.
[18]. Timakova A. et al. Computer Vision-Assisted Spatial Analysis of Mitoses and Vasculature in Lung Cancer. Journal of Clinical Medicine, 2025, vol. 14, no. 21, p. 7526.
[19]. Асатурова А.В., Трегубова А.В., Бадлаева А.С., Рогожина А.С., Ананьев В.В., Карпулевич Е.А., Рогожин Д.В. Возможности цифровой цитопатологии, проблемы ее внедрения в практику и пути их решения. Журнал Современные проблемы науки и образования, 2025, № 3.

Информация об авторах / Information about authors

Владислав Валерьевич АНАНЬЕВ – научный сотрудник 30-го центра ИСП РАН. Сфера научных интересов: применение нейронных сетей и методов обработки изображений к данным биомедицинского домена, разработка и оптимизация вычислительных конвейеров для предобработки изображений, применение языковых моделей и RAG-систем для обработки и систематизации медицинских данных.

Vladislav Valeryevich ANANEV – research officer at the 30th center of ISP RAS. Research interests: application of neural networks and image processing methods to biomedical data,

development and optimization of computational pipelines for image preprocessing, application of language models and RAG systems for processing and systematizing medical data.

Эдуард Владимирович ЕФИМОВ – лаборант 30-го центра ИСП РАН. Сфера научных интересов: машинное обучение, компьютерное зрение, LLM, прикладная математика, анализ данных.

Eduard Vladimirovich EFIMOV – laboratory assistant of the 30th center of ISP RAS. Research interests: machine learning, computer vision, LLMs, applied mathematics, data analysis.

Егор Сергеевич ЗЕМНУХОВ – младший научный сотрудник 30-го центра ИСП РАН. Сфера научных интересов: машинное обучение и компьютерное зрение, применяемые в области анализа медицинских изображений, разработка инструментов для работы с биомедицинскими данными.

Egor Sergeyevich ZEMNUKHOV – research assistant at the 30th center of ISP RAS. Research interests: machine learning and computer vision applied to medical image analysis, development of tools for working with biomedical data.

Анна Алексеевна ТИМАКОВА – научный сотрудник Сеченовского Университета. Сфера научных интересов: анализ биомедицинских изображений при помощи компьютерного зрения.

Anna Alekseyevna TIMAKOVA is a research officer at Sechenov University. Research interests: biomedical image analysis using computer vision.

Евгений Андреевич КАРПУЛЕВИЧ – кандидат физико-математических наук, заведующий 30-го центра ИСП РАН. Сфера научных интересов: применение алгоритмов анализа данных к биомедицинскому домену, разработка масштабируемых программных конвейеров анализа биомедицинских данных.

Evgeny Andreyevich KARPULEVICH – Cand. Sci. (Phys.-Math.), head of the 30th center of ISP RAS. Research interests: application of data analysis algorithms to the biomedical domain, development of scalable software pipelines for biomedical data analysis.