МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра компьютерных технологий и систем

Борычев
Денис Валерьевич

ГЕНЕРАЦИЯ КАРТ ПО СПУТНИКОВЫМ СНИМКАМ

Дипломная работа

Научный руководитель:
старший преподаватель кафедры
компьютерных технологий и систем
Соловей С.С.

Допущен к защите
«____» _____ 2025 г.

Зав. кафедрой компьютерных технологий и систем
кандидат физ.-мат. наук, доктор педагогических наук
профессор В. В. Казаченок

Минск, 2025

РЕФЕРАТ

Дипломная работа, 57 страниц, 12 рисунков, 1 таблица, 0 приложений, 12 источников.

Ключевые слова: ГЕНЕРАЦИЯ КАРТ, СПУТНИКОВЫЕ СНИМКИ, ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫЕ СЕТИ, VISION TRANSFORMER, PIX2PIX, ГЛУБОКОЕ ОБУЧЕНИЕ, КАРТОГРАФИЯ, ГЕОИНФОРМАЦИОННЫЕ СИСТЕМЫ.

Объект исследования – методы генерации карт по спутниковым изображениям с использованием архитектур глубокого обучения.

Предмет исследования – процесс генерации карт по спутниковым снимкам с использованием моделей глубокого обучения, включая архитектуры GAN и Vision Transformer.

Цель исследования – провести сравнительный анализ эффективности традиционной модели pix2pix и экспериментальной модели GAN с генератором на основе Vision Transformer для задач генерации карт по спутниковым снимкам.

Методы исследования – экспериментальное исследование, сравнительный анализ архитектур нейронных сетей, количественная и качественная оценка результатов генерации.

Полученные результаты и их новизна: подтверждена перспективность применения трансформерных архитектур в задачах генерации изображений; продемонстрированы преимущества механизмов внимания для моделирования глобальных пространственных зависимостей; обоснована эффективность гибридных подходов, сочетающих Vision Transformer с генеративными моделями.

Достоверность материалов и результатов дипломной работы: используемые материалы и результаты дипломной работы являются достоверными. Работа выполнена самостоятельно.

Область возможного практического применения – геоинформатика, автоматизированное картографирование, обработка спутниковых данных, геоинформационные системы.

РЭФЕРАТ

Дыпломная праца, 57 старонак, 12 малюнкаў, 1 табліца, 0 дадаткаў, 12 крыніц.

Ключавыя словы: ГЕНЕРАЦЫЯ КАРТ, СПУТНІКОВЫЯ ЗДЫМКІ, ГЕНЕРАТЫЎНА-СОСЦЯЗАЛЬНЫЯ СЕТКІ, VISION TRANSFORMER, PIX2PIX, ГЛЫБОКАЕ НАВУЧАННЕ, КАРТАГРАФАВАННЕ, ГЕАІНФАРМАЦЫЙНЫЯ СІСТЭМЫ.

Аб'ект даследавання – метады генерацыі карт па спадарожнікавых выявах з выкарыстаннем архітэктур глыбокага навучання.

Прадмет даследавання – працэс генерацыі карт па спадарожнікавых здымках з выкарыстаннем мадэляў глыбокага навучання, уключаючы архітэктуры GAN і Vision Transformer.

Мэта даследавання – правесці параўнальны аналіз эфектыўнасці традыцыйнай мадэлі pix2pix і эксперыментальнай мадэлі GAN з генератарам на аснове Vision Transformer для задач генерацыі карт па спадарожнікавых здымках.

Методы даследавання – эксперыментальнае даследаванне, параўнальны аналіз архітэктур нейронных сетак, колькасная і якасная ацэнка вынікаў генерацыі.

Атрыманыя вынікі і іх навізна: пацверджана перспектыўнасць прымянення трансфармерных архітэктур у задачах генерацыі выяў; прадэманстраваны перавагі механізмаў увагі для мадэлявання глабальных прасторавых залежнасцей; абгрунтавана эфектыўнасць гібрыдных падыходаў, якія спалучаюць Vision Transformer з генератыўнымі мадэлямі.

Дакладнасць матэрыялаў і вынікаў дыпломнай працы: выкарыстаныя матэрыялы і вынікі дыпломнай працы з'яўляюцца дастаўернымі. Работа выканана самастойна.

Вобласць магчымага практычнага прымянення – геаінфарматыка, аўтаматызаванае картографаванне, апрацоўка спадарожнікавых дадзеных, геаінфармацыйныя сістэмы.

ABSTRACT

Diploma work, 57 pages, 12 figures, 1 table, 0 appendixes, 12 references.

Keywords: MAP GENERATION, SATELLITE IMAGES, GENERATIVE ADVERSARIAL NETWORKS, VISION TRANSFORMER, PIX2PIX, DEEP LEARNING, CARTOGRAPHY, GEOGRAPHIC INFORMATION SYSTEMS.

The object of the research – methods for generating maps from satellite imagery using deep learning architectures.

The subject of the research – the process of map generation from satellite images using deep learning models, including GAN and Vision Transformer architectures.

The aim of the research – to conduct a comparative analysis of the effectiveness of the traditional pix2pix model and an experimental GAN model with a Vision Transformer-based generator for the task of map generation from satellite imagery.

Research methods – experimental study, comparative analysis of neural network architectures, quantitative and qualitative evaluation of generated outputs.

The results of the work and their novelty: the prospects of applying transformer-based architectures in image generation tasks were confirmed; advantages of attention mechanisms in modeling global spatial dependencies were demonstrated; the effectiveness of hybrid approaches combining Vision Transformer with generative models was substantiated.

Authenticity of the materials and results of the diploma work: the materials and results presented in this diploma work are reliable. The work was completed independently.

Recommendations on the usage: geoinformatics, automated cartography, satellite data processing, geographic information systems.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Развитие дистанционного зондирования Земли и рост объёмов данных, получаемых со спутников, открывают новые горизонты для решения широкого спектра геоинформационных задач. В условиях постоянного расширения сфер применения картографических материалов — от мониторинга природных ландшафтов и оценки состояния экосистем до планирования городской инфраструктуры и управления ресурсами — качественные и точные карты становятся критически важным инструментом. Однако формирование карт из спутниковых снимков является сложной задачей, требующей комплексной обработки больших объёмов данных, выделения ключевых объектов и структур, а также обеспечения высокого уровня детализации и реалистичности.

Одним из центральных направлений в данной области выступает генерация карт со спутниковых изображений, которая в последние годы дополняется и совершенствуется методами машинного обучения и глубокой сегментации. Традиционные алгоритмические подходы, основанные на пороговой фильтрации, морфологических операциях и ручной разметке, часто оказываются недостаточно точными, трудоёмкими и слабо масштабируемыми при больших объёмах данных. Недостатки классических методов особенно ощутимы при необходимости учитывать неоднородность местности, сложный рельеф и разнообразие земного покрова.

Современные методы глубокого обучения предлагают новые пути решения данных проблем. В частности, генеративные состязательные сети (GAN, Generative Adversarial Networks) и их условные варианты позволяют создавать реалистичные изображения и осуществлять перевод данных из одного домена в другой. Эти подходы особенно эффективны для формирования высокодетализированных карт на основе спутниковых снимков. Среди наиболее перспективных архитектур можно выделить модель *pix2pix*, которая успешно используется в задачах перевода изображений, включая задачи картографирования земной поверхности.

Параллельно с развитием традиционных свёрточных архитектур, в области компьютерного зрения активно исследуются алтернативные подходы, основанные на механизмах внимания (attention). Vision Transformer (ViT) представляет собой революционную архитектуру, которая адаптирует принципы трансформеров для работы с изображениями. Применение ViT в качестве ге-

нератора в составе генеративно-состязательных сетей открывает новые возможности для решения задач генерации карт, потенциально превосходя существующие подходы в точности и качестве результатов.

Использование высокоразрешённых спутниковых данных в сочетании с нейросетевыми моделями даёт возможность автоматизировать процесс генерации карт и повысить точность результатов. Это позволяет оперативно получать детализированные картографические материалы, применимые в исследованиях динамики ландшафта, мониторинге изменений окружающей среды, оптимизации землепользования, стратегическом планировании инфраструктурных проектов и решении множества других прикладных задач.

Таким образом, настоящее исследование сосредоточено на сравнительном анализе применения двух различных архитектур глубокого обучения для генерации карт со спутниковых снимков: классической модели *pix2pix* и экспериментальной модели GAN с генератором на основе Vision Transformer. В работе будут рассмотрены существующие подходы, их преимущества и ограничения, обоснован выбор каждой из архитектур и продемонстрирована их адаптация к поставленной задаче. Особое внимание уделяется анализу производительности, качества результатов и вычислительной эффективности каждого подхода. Ожидается, что полученные результаты расширят теоретическую и практическую базу в области геоинформатики и автоматизации картографирования, а также позволят оценить перспективы использования современных архитектур, основанных на механизмах внимания, в задачах генерации геопространственных данных.

ГЛАВА 1
ПРОБЛЕМА ГЕНЕРАЦИИ КАРТ СО СПУТНИКОВЫХ СНИМКОВ

1.1 Применение генерации спутниковых изображений

Генерация карт по спутниковым изображениям представляет собой одно из наиболее динамично развивающихся направлений в современной геоинформатике и дистанционном зондировании Земли. Данная область приобретает особую важность в контексте растущих потребностей общества в оперативном получении точной пространственной информации для решения широкого спектра научных, практических и управленческих задач.

Рисунок 1.1 — Пример спутникового изображения для генерации карты

Современные спутники высокого разрешения предоставляют исследователям уникальные возможности для получения детализированных изображений земной поверхности с пространственным разрешением от нескольких метров до десятков сантиметров(рисунок 1.1). Эти данные открывают новые горизонты для создания высокоточных картографических материалов, включающих топографические карты, карты землепользования, инфраструктурные схемы и специализированные тематические карты.

Спектр применения автоматически генерируемых карт чрезвычайно широк. В области экологического мониторинга такие карты используются для отслеживания изменений растительного покрова, мониторинга лесных массивов, выявления процессов деградации земель и оценки биоразнообразия. Они позволяют оперативно выявлять нарушения экологического баланса, отсле-

живать динамику природных процессов и прогнозировать возможные экологические риски.

В сфере городского планирования автоматизированная генерация карт обеспечивает возможность детального картирования городской инфраструктуры, включая здания, дороги, зеленые зоны и коммуникации. Это критически важно для планирования развития городов, оптимизации транспортных сетей, управления водными ресурсами и создания эффективных систем городского управления. Особую ценность представляет возможность оперативного обновления картографической информации для отражения быстро изменяющейся городской среды.

Сельскохозяйственные приложения и программы включают создание карт посевных площадей, мониторинг состояния сельскохозяйственных культур, оценку урожайности и планирование агротехнических мероприятий. Прецизионное земледелие, основанное на детальных картах полей, позволяет оптимизировать использование удобрений, пестицидов и водных ресурсов, что приводит к повышению эффективности сельскохозяйственного производства и снижению экологической нагрузки.

В области управления природными ресурсами автоматически генерируемые карты используются для инвентаризации лесных ресурсов, мониторинга водных объектов, планирования добычи полезных ископаемых и управления особо охраняемыми природными территориями. Они обеспечивают научную основу для принятия решений в области природопользования и помогают балансировать экономические интересы с экологическими требованиями.

Особое значение имеет применение технологий генерации карт в системах реагирования на чрезвычайные ситуации. При возникновении стихийных бедствий, таких как наводнения, пожары, землетрясения или техногенные катастрофы, оперативное создание карт пострадавших территорий критически важно для координации спасательных операций, планирования эвакуации и оценки ущерба. Автоматизированные системы генерации карт позволяют значительно сократить время получения необходимой пространственной информации, что может спасти множество жизней.

Научные исследования также существенно выигрывают от применения современных методов генерации карт. Климатологи используют такие карты для изучения изменений климата и их локальных проявлений. Геологи применяют их для картирования геологических структур и изучения геодинамических процессов. Географы исследуют с их помощью пространственные

закономерности природных и социально-экономических явлений.

Образовательные приложения включают создание картографических материалов для обучения студентов и школьников основам географии, экологии и пространственного анализа. Интерактивные карты, созданные на основе спутниковых данных, позволяют наглядно демонстрировать пространственные процессы и явления, что значительно повышает эффективность образовательного процесса.

Таким образом, генерация карт по спутниковым изображениям представляет собой фундаментальную технологию, обеспечивающую научную и информационную основу для решения множества актуальных задач современного общества. Развитие данного направления непосредственно связано с прогрессом в области обработки больших данных, методов машинного обучения и вычислительных технологий.

1.2 Традиционные методы и современные технологии обработки спутниковых данных

История развития методов обработки спутниковых изображений для картографических целей насчитывает несколько десятилетий и характеризуется последовательным переходом от простых алгоритмических подходов к сложным интеллектуальным системам. Понимание этой эволюции критически важно для осознания современных вызовов и перспектив развития области.

Традиционные методы обработки спутниковых данных основывались на принципах классического анализа изображений и включали различные техники сегментации, классификации и фильтрации. Пороговая сегментация, представляющая собой один из первых автоматизированных подходов, заключалась в разделении изображения на области на основе пороговых значений яркости или других спектральных характеристик. Несмотря на простоту реализации, этот метод оказывался неэффективным при работе с изображениями, содержащими объекты со схожими спектральными характеристиками или при наличии существенного шума.

Морфологические операции, такие как эрозия, дилатация, открытие и закрытие, применялись для улучшения результатов сегментации и удаления артефактов. Эти методы позволяли устранять мелкие шумовые элементы и заполнять небольшие пробелы в выделенных объектах. Однако они требова-

ли тщательной настройки параметров и часто приводили к потере важных деталей изображения.

Методы кластеризации, включая k-means и ISODATA (Iterative Self-Organizing Data Analysis Technique), использовались для группировки пикселей с похожими спектральными характеристиками. Эти подходы позволяли выявлять различные типы земного покрова без предварительного знания количества классов, что делало их привлекательными для исследовательских задач. Тем не менее, результаты кластеризации часто требовали значительной постобработки и интерпретации экспертом.

Существенным недостатком традиционных методов являлась их ограниченная способность учитывать пространственный контекст. Большинство алгоритмов анализировали пиксели независимо друг от друга, игнорируя важную информацию о пространственных отношениях между объектами. Это приводило к фрагментированным результатам сегментации и необходимости дополнительной постобработки для получения связных объектов.

С развитием методов машинного обучения в обработку спутниковых данных стали внедряться более совершенные алгоритмы классификации. Методы опорных векторов (SVM) продемонстрировали высокую эффективность в задачах классификации спектральных характеристик пикселей. Случайные леса (Random Forest) обеспечивали хорошую производительность при работе с многомерными данными и позволяли оценивать важность различных спектральных каналов.

Однако даже эти методы имели существенные ограничения. Они требовали ручного конструирования признаков, что было трудоемким процессом и требовало глубокой экспертизы в области дистанционного зондирования. Кроме того, получаемые модели часто плохо генерализовались на данные, отличающиеся по условиям съемки или географическому положению.

Современные программные решения представляют собой комплексные системы, интегрирующие различные методы обработки и анализа пространственных данных. QGIS и ArcGIS являются наиболее популярными геоинформационными системами общего назначения, предоставляющими широкий набор инструментов для работы со спутниковыми изображениями (рисунок 1.2). Эти системы включают модули для предварительной обработки данных, различные алгоритмы классификации и сегментации, а также средства визуализации и анализа результатов.

Специализированные пакеты, такие как ENVI и ERDAS IMAGINE, пред-

назначены специально для работы с данными дистанционного зондирования. Они предоставляют продвинутые инструменты для калибровки спутниковых изображений, атмосферной коррекции, спектрального анализа и тематического картирования. Эти системы особенно популярны среди исследователей и специалистов, работающих с многоспектральными и гиперспектральными данными.

Google Earth Engine представляет собой революционную облачную платформу, предоставляющую доступ к петабайтам спутниковых данных и мощным вычислительным ресурсам для их обработки. Платформа позволяет выполнять масштабные геопространственные анализы, включающие временные ряды спутниковых изображений, что открывает новые возможности для мониторинга изменений земной поверхности.

С появлением методов глубокого обучения область обработки спутниковых данных претерпела кардинальные изменения. Сверточные нейронные сети продемонстрировали способность автоматически извлекать иерархические признаки из изображений, что привело к значительному улучшению качества сегментации и классификации. Архитектуры, такие как U-Net, SegNet и DeepLab, стали стандартом в задачах семантической сегментации спутниковых изображений.

Рисунок 1.2 — Пример работы программы QGIS

Современные тенденции в области обработки спутниковых данных включают использование глубоких нейронных сетей для решения сложных задач объектно-ориентированного анализа изображений. Методы instance segmentation позволяют не только классифицировать пиксели по типам объектов, но и выделять отдельные экземпляры объектов. Это особенно важно для картирования зданий, транспортных средств и других дискретных объектов на

спутниковых изображениях.

Другим перспективным направлением является использование временных рядов спутниковых данных для анализа динамики изменений земной поверхности. Рекуррентные нейронные сети и трансформеры приспособлены для работы с последовательностями изображений, что позволяет выявлять тренды и аномалии в развитии различных процессов.

Интеграция различных источников данных, включая спутниковые изображения различного разрешения, LIDAR данные и открытые географические базы данных, открывает новые возможности для создания комплексных картографических продуктов. Методы multi-modal learning позволяют эффективно комбинировать информацию из различных источников для повышения точности и полноты картирования.

Таким образом, современное состояние области обработки спутниковых изображений для автоматизированного картографирования характеризуется активным поиском новых архитектур и методов, способных преодолеть ограничения традиционных подходов и обеспечить более высокое качество автоматизированной генерации карт по спутниковым данным.

1.3 Современные методы глубокого обучения в картографии

Революция в области обработки спутниковых изображений, связанная с применением методов глубокого обучения, началась в середине 2010-х годов и продолжает активно развиваться. Генеративные состязательные сети (GAN) представляют собой одно из наиболее перспективных направлений в этой области, предлагая принципиально новый подход к решению задач генерации картографических материалов.

Концепция GAN основана на состязательном обучении двух нейронных сетей: генератора и дискриминатора. Генератор стремится создавать изображения, неотличимые от реальных, в то время как дискриминатор пытается различить сгенерированные и настоящие изображения. Этот процесс можно сравнить с игрой между фальшивомонетчиком (генератором) и экспертом (дискриминатором), где каждая сторона постоянно совершенствует свои навыки в ответ на улучшения противника.

В контексте картографических задач GAN демонстрируют исключительную способность обучаться сложным пространственным паттернам и воспро-

изводить их в создаваемых картах. Они способны захватывать не только локальные особенности объектов, но и глобальную структуру изображения, что критически важно для создания согласованных и реалистичных картографических представлений.

Особую ценность для задач генерации карт по спутниковым снимкам представляют условные генеративные модели (Conditional GAN). В отличие от обычных GAN, которые генерируют изображения из случайного шума, условные модели создают выходные изображения на основе входных данных определенного типа. Это позволяет осуществлять направленное преобразование спутниковых изображений в картографические представления с сохранением пространственного соответствия и семантической корректности.

Архитектура pix2pix представляет собой один из наиболее успешных примеров применения условных GAN для задач image-to-image translation. Эта модель специально разработана для выполнения парных преобразований изображений, что делает ее идеально подходящей для генерации карт по спутниковым снимкам. Pix2pix сочетает преимущества генеративного моделирования с возможностью точного контроля процесса генерации через входные изображения.

Ключевой особенностью pix2pix является использование комбинированной функции потерь, включающей как состязательную потерю (adversarial loss), так и L1 потерю. Состязательная потеря обеспечивает реалистичность генерируемых изображений, заставляя генератор создавать карты, которые дискриминатор не может отличить от настоящих. L1 потеря, в свою очередь, обеспечивает пиксельное соответствие между входным изображением и выходной картой, что критически важно для сохранения точности пространственного позиционирования объектов.

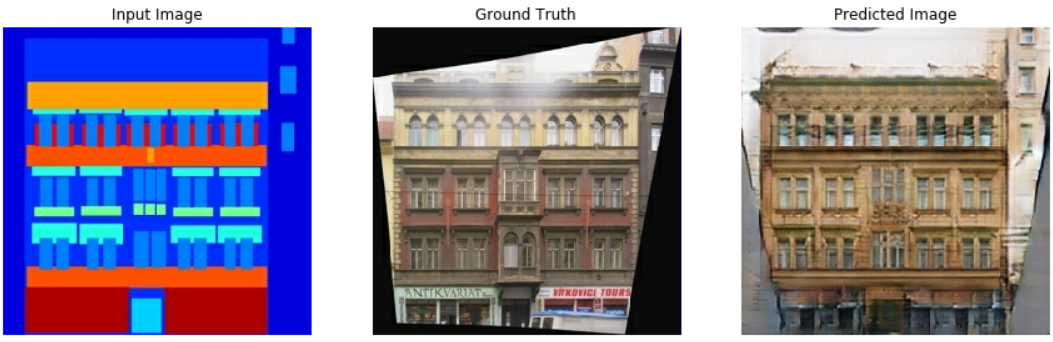

Рисунок 1.3 — Пример работы модели pix2pix для генерации карт

Применение pix2pix в картографических задачах демонстрирует впечат-

ляющие результаты в самых различных сценариях (рисунок 1.3). Модель успешно справляется с генерацией различных типов карт, включая топографические карты, карты землепользования, дорожные сети и изображения городской инфраструктуры. Особенно важной является способность модели адаптироваться к различным типам входных данных, включая многоспектральные изображения, данные радиолокационной съемки и комбинированные датасеты.

Однако, несмотря на высокую эффективность, pix2pix и другие модели на основе сверточных архитектур имеют определенные ограничения. Основное из них связано с локальной природой сверточных операций, которые ограничивают способность модели моделировать дальние зависимости в изображении. Это может приводить к проблемам с воспроизведением глобальной структуры карт, особенно при работе с изображениями большого размера или содержащими объекты значительного протяжения.

Развитие области генеративного моделирования характеризуется постоянным поиском новых архитектур и методов обучения, направленных на преодоление текущих ограничений. Исследователи работают над улучшением стабильности обучения GAN, повышением качества генерации мелких деталей и развитием методов контроля процесса генерации. Эти усилия открывают новые перспективы для создания еще более точных и детализированных картографических материалов.

1.4 Перспективные направления и новые архитектуры

В последние годы в области компьютерного зрения происходит значительный сдвиг от традиционных сверточных архитектур к архитектурам на основе механизмов внимания, известным как трансформеры. Этот переход обусловлен фундаментальными ограничениями сверточных операций и потенциальными преимуществами механизмов самовнимания (self-attention) для моделирования сложных пространственных отношений.

Традиционные сверточные сети ограничены локальными рецептивными полями, что затрудняет моделирование дальних зависимостей в изображениях. Для спутниковых снимков, которые часто содержат объекты значительного протяжения, такие как реки, дороги или горные цепи, эти ограничения могут существенно влиять на качество анализа. Механизмы внимания позволяют напрямую моделировать взаимодействия между любыми частями

изображения, независимо от их пространственного расстояния.

Vision Transformer (ViT) адаптировал принципы трансформеров, первоначально разработанных для обработки естественного языка, для задач компьютерного зрения. Основная идея ViT заключается в разбиении изображения на патчи фиксированного размера, которые затем обрабатываются как последовательность токенов аналогично словам в тексте. Каждый патч преобразуется в вектор признаков с добавлением позиционного кодирования, после чего последовательность патчей передается на вход трансформера (рисунок 1.4).

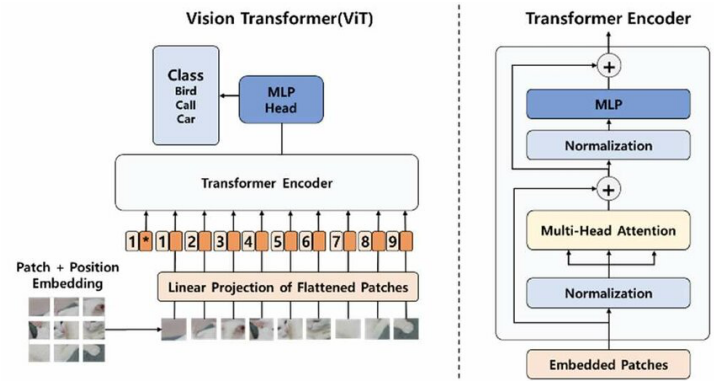

Рисунок 1.4 — Концептуальная схема Vision Transformer

Для генерации карт по спутниковым снимкам ViT предлагает несколько потенциальных преимуществ. Прежде всего, глобальная природа механизмов внимания позволяет модели эффективно захватывать дальние пространственные зависимости. Это особенно важно для понимания связности объектов на карте, таких как дорожные сети или речные системы, которые могут простираться через всё изображение.

Масштабируемость представляет еще одно важное преимущество ViT. В отличие от сверточных сетей, которые требуют адаптации для работы с изображениями различного размера, трансформеры могут более естественно обрабатывать последовательности патчей переменной длины. Это позволяет использовать одну и ту же архитектуру для обработки спутниковых изображений различного разрешения без значительных модификаций.

Эффективность обучения трансформеров при наличии больших объемов данных также представляет интерес для картографических приложений. Спутниковые данные характеризуются высокой доступностью и постоянным обновлением, что создает благоприятные условия для обучения больших моделей. Исследования показывают, что трансформеры могут демонстрировать

более стабильное обучение и лучшую генерализацию по сравнению с сверточными архитектурами при использовании больших датасетов.

Интеграция ViT в генеративные модели для картографических задач представляет активно развивающееся направление исследований. Существуют различные подходы к такой интеграции, от простой замены сверточного генератора на трансформерный до разработки гибридных архитектур, сочетающих преимущества обоих подходов. Каждый из этих подходов имеет свои преимущества и ограничения, требующие тщательного исследования.

Потенциальные вызовы применения ViT в генеративных моделях включают высокие вычислительные требования, особенно для изображений высокого разрешения, и необходимость больших объемов данных для эффективного обучения. Кроме того, интеграция трансформерной архитектуры в состязательное обучение может потребовать разработки новых стратегий обучения и функций потерь.

Несмотря на эти вызовы, перспективы применения ViT в задачах генерации карт выглядят многообещающими. Возможность эффективного моделирования глобальных пространственных отношений может привести к созданию более точных и согласованных картографических представлений, особенно для крупномасштабных объектов и сложных пространственных структур.

1.5 Постановка задачи и методология исследования

В данной работе проводится сравнительный анализ двух перспективных подходов к генерации карт по спутниковым снимкам. На основе анализа современного состояния области были выбраны следующие модели как наиболее перспективные и научно обоснованные для исследования.

Модель pix2pix была выбрана как представитель зрелого и хорошо изученного подхода на основе генеративных состязательных сетей с сверточной архитектурой. Данный выбор обусловлен несколькими факторами. Во-первых, pix2pix продемонстрировал стабильную эффективность в широком спектре задач image-to-image translation, включая картографические приложения. Во-вторых, модель имеет относительно простую архитектуру и хорошо документированные принципы работы, что обеспечивает воспроизводимость результатов и возможность объективного анализа. В-третьих, существует обширная база исследований, посвященных применению pix2pix в различных доменах, что позволяет провести сравнение с предыдущими ра-

ботами.

Экспериментальная модель GAN с генератором на основе Vision Transformer была выбрана как представитель инновационного направления, интегрирующего достижения в области архитектур внимания с генеративным моделированием. Этот выбор мотивирован потенциальными преимуществами трансформеров в моделировании глобальных пространственных зависимостей, что критически важно для понимания структуры спутниковых изображений. Кроме того, применение ViT в генеративных задачах представляет собой относительно новое направление исследований, что делает сравнение с классическими подходами особенно ценным с научной точки зрения.

Основная цель исследования состоит в проведении всестороннего сравнительного анализа этих двух подходов по нескольким ключевым аспектам. Качество генерации оценивается через анализ точности и детализации создаваемых карт, способности моделей воспроизводить сложные пространственные структуры и сохранять семантическую корректность объектов. Вычислительная эффективность анализируется с точки зрения времени обучения, скорости инференса, требований к вычислительным ресурсам и масштабируемости для работы с изображениями различного размера.

Практическая применимость исследуется через определение оптимальных сценариев использования каждого подхода, анализ их ограничений и возможностей адаптации для различных типов картографических задач. Особое внимание уделяется вопросам стабильности обучения, чувствительности к гиперпараметрам и способности моделей генерализоваться на данные из различных географических регионов.

Методология исследования построена на принципах научной строгости и объективности. Для обеспечения справедливого сравнения используется единый тщательно подготовленный набор данных с идентичным разделением на обучающую, валидационную и тестовую выборки. Набор включает разнообразные типы спутниковых изображений и соответствующих им карт, представляющих различные географические регионы и типы ландшафтов.

Метрики оценки качества выбраны таким образом, чтобы обеспечить комплексную оценку различных аспектов генерации карт. Количественные метрики включают как традиционные показатели качества изображений, так и специализированные картографические метрики, оценивающие сохранение топологических отношений и точность воспроизведения пространственных структур. Качественная оценка проводится через экспертный анализ сгене-

рированных карт с привлечением специалистов в области картографии и дистанционного зондирования.

Анализ вычислительных характеристик осуществляется в контролируемых условиях с использованием идентичного аппаратного обеспечения и программной среды. Проводится детальное профилирование производительности моделей, включая анализ потребления памяти, вычислительной сложности и временных характеристик различных этапов обучения и инференса.

Для обеспечения статистической значимости результатов эксперименты проводятся многократно с различными инициализациями параметров модели. Результаты анализируются с применением соответствующих статистических методов для определения значимости различий между моделями и оценки доверительных интервалов для основных метрик.

Детальное описание каждого подхода, включая архитектурные особенности, процесс обучения и результаты экспериментов, будет представлено в последующих главах. Во второй главе будет подробно рассмотрена теоретическая основа и практическая реализация модели pix2pix. Третья глава посвящена экспериментальной модели с интеграцией Vision Transformer. Четвертая глава представит результаты сравнительного анализа и их интерпретацию.

1.6 Выводы к главе 1

Данная глава представила всесторонний анализ проблемы генерации карт по спутниковым снимкам, охватывающий как исторический контекст, так и современное состояние области. Эволюция методов от традиционных алгоритмических подходов к современным архитектурам глубокого обучения демонстрирует значительный прогресс в области автоматизированного картографирования.

Традиционные методы, основанные на пороговой сегментации, морфологических операциях и классических алгоритмах машинного обучения, заложили фундамент для развития области, но показали существенные ограничения в точности, масштабируемости и способности обрабатывать сложные пространственные структуры. Переход к методам глубокого обучения, особенно к генеративным состязательным сетям, открыл новые возможности для создания высококачественных картографических материалов.

Современные подходы, представленные моделью pix2pix, демонстрируют высокую эффективность в задачах генерации карт, сочетая способности гене-

ративного моделирования с точным контролем процесса создания картографических представлений. Однако ограничения сверточных архитектур в моделировании дальних пространственных зависимостей создают потребность в более продвинутых решениях.

Появление архитектур на основе трансформеров, особенно Vision Transformer, представляет новое перспективное направление развития. Способность этих архитектур эффективно моделировать глобальные зависимости в изображениях может привести к значительному улучшению качества генерации карт, особенно для объектов большого протяжения и сложных пространственных структур.

Выбор модели pix2pix и экспериментальной модели GAN с ViT-генератором для сравнительного анализа обоснован их представительностью для двух различных парадигм в области генерации изображений: проверенного сверточного подхода и инновационного трансформерного направления. Сравнение этих подходов позволит получить важные результаты о применимости современных архитектур внимания для картографических задач и оценить баланс между качеством генерации и вычислительной эффективностью.

Методология исследования, основанная на принципах научной строгости и объективности, обеспечит получение надежных и воспроизводимых результатов. Использование единого датасета, стандартизованных метрик и контролируемых экспериментальных условий позволит провести честное сравнение двух подходов и сформулировать обоснованные выводы об их относительных преимуществах и недостатках.

Результаты данного исследования внесут вклад в понимание потенциала современных архитектур машинного обучения для решения картографических задач и предоставят ценную информацию для исследователей и практиков, работающих в области автоматизированной обработки спутниковых данных.

ГЛАВА 2

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ И РЕАЛИЗАЦИЯ МОДЕЛИ PIX2PIX

2.1 Генеративные модели: вариационные автокодировщики

До появления генеративно-состязательных нейронных сетей, для задач генерации данных широко использовались вариационные автокодировщики (Variational Autoencoders, VAE). Эти модели представляют собой важную веху в развитии генеративных алгоритмов и заложили теоретические основы для понимания проблем генерации изображений, которые впоследствии были решены с помощью GAN.

Вариационные автокодировщики основаны на принципах вариационной байесовской инференции и представляют собой вероятностную модель, которая обучается восстанавливать входные данные через скрытое представление. Ключевая идея VAE заключается в предположении, что наблюдаемые данные x генерируются из скрытых переменных z согласно некоторому распределению $p(x|z)$.

Математически VAE можно описать следующим образом. Пусть $p(z)$ - априорное распределение скрытых переменных (обычно стандартное нормальное распределение), $p(x|z)$ - вероятность генерации данных по скрытым переменным, а $q(z|x)$ - приближенное апостериорное распределение, параметризованное энкодером. Цель VAE состоит в максимизации нижней границы правдоподобия (Evidence Lower BOund, ELBO):

$$\mathcal{L}_{VAE} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)), \tag{2.1}$$

VAE состоят из трех основных компонентов, каждый из которых играет критическую роль в процессе генерации. Энкодер $q_\phi(z|x)$ представляет собой нейронную сеть, которая принимает входные данные x и выводит параметры распределения в скрытом пространстве. Важно отметить, что энкодер не просто создает детерминистическое представление, а генерирует параметры вероятностного распределения - обычно среднее $\mu(x)$ и логарифм дисперсии $\log \sigma^2(x)$ нормального распределения. Это позволяет модели моделировать

неопределенность в скрытых представлениях.

Скрытое пространство в VAE имеет фундаментальное значение. Это многомерное пространство, где каждая точка представляет возможное скрытое представление входных данных. Ключевое свойство скрытого пространства VAE заключается в том, что оно структурировано таким образом, что похожие данные отображаются в близкие точки, а семплирование из этого пространства должно давать правдоподобные данные. Размерность скрытого пространства обычно значительно меньше размерности входных данных, что заставляет модель изучать компактное и содержательное представление.

Декодер $p_\theta(x|z)$ выполняет обратное преобразование, принимая точку из скрытого пространства и генерируя соответствующие данные. В случае генерации изображений декодер обычно реализуется с помощью транспонированных сверток или других операций апсэмплинга, которые постепенно увеличивают пространственное разрешение от скрытого представления до полноразмерного изображения (рисунок 2.1).

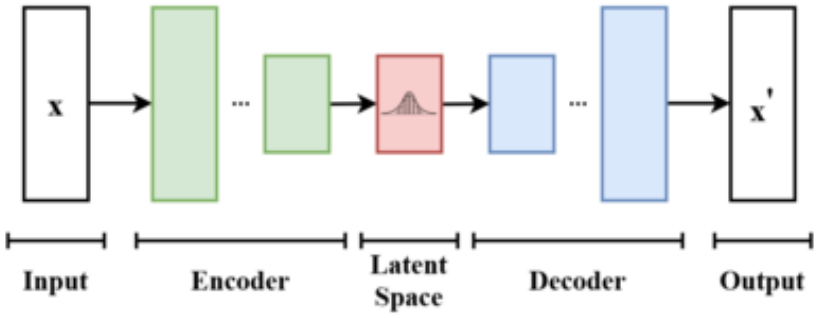

Рисунок 2.1 — Схема вариационного автокодировщика

Процесс обучения VAE включает в себя так называемый "трюк репараметризации"(reparameterization trick), предложенный Кингмой и Веллингом. Поскольку семплирование из распределения $q(z|x)$ является недифференцируемой операцией, невозможно напрямую вычислить градиенты. Трюк репараметризации решает эту проблему путем перезаписи семплирования как $z = \mu + \sigma \odot \epsilon$, где $\epsilon \sim \mathcal{N}(0, I)$ - вспомогательная случайная переменная, $\odot$ обозначает поэлементное произведение. Это позволяет переместить стохастичность в независимую переменную и сделать процесс дифференцируемым.

Функция потерь VAE состоит из двух основных компонентов. Первый компонент - это потеря реконструкции, обычно измеряемая как среднеквадратичная ошибка (MSE) или биномиальная кросс-энтропия между входными

и восстановленными данными: $\mathcal{L}_{recon} = \mathbb{E}_{q(z|x)}[\log p(x|z)]$. Второй компонент - это регуляризационная потеря KL-дивергенции: $\mathcal{L}_{KL} = D_{KL}(q(z|x)||p(z))$, которая заставляет выученное апостериорное распределение быть близким к априорному.

KL-дивергенция играет критически важную роль в обучении VAE и решает несколько ключевых проблем. Во-первых, она обеспечивает регуляризацию скрытого пространства, предотвращая "коллапс"энкодера в детерминистические представления. Без KL-регуляризации энкодер мог бы просто игнорировать стохастическую природу скрытых переменных и превратиться в обычный автокодировщик. Во-вторых, KL-дивергенция способствует созданию структурированного и интерполируемого скрытого пространства, где семплирование из любой точки должно давать правдоподобные данные.

Однако, несмотря на элегантность теоретических основ, VAE сталкиваются с серьезными практическими проблемами при генерации изображений. Наиболее заметной из них является проблема размытости сгенерированных изображений. Эта проблема имеет несколько причин.

Во-первых, использование MSE в качестве потери реконструкции приводит к усреднению. Когда несколько разных изображений могут соответствовать одному скрытому представлению, MSE заставляет декодер генерировать среднее значение всех возможных выходов, что неизбежно приводит к размытости. Например, если скрытое представление может соответствовать как изображению с дорогой слева, так и с дорогой справа, MSE заставит модель генерировать размытое изображение с "дорогой посередине".

Во-вторых, Posterior Collapse - явление, при котором энкодер игнорирует входные данные и всегда выводит распределение, близкое к априорному. В этом случае вся информация о входных данных теряется, и декодер вынужден генерировать "средние"изображения для всех входов. Это особенно проблематично для задач, требующих точного воспроизведения деталей, таких как генерация карт.

Третьей проблемой является фундаментальный компромисс между качеством реконструкции и регуляризацией скрытого пространства. Увеличение веса KL-дивергенции улучшает структуру скрытого пространства, но ухудшает способность точно воспроизводить входные данные. Уменьшение этого веса приводит к лучшей реконструкции, но может разрушить полезную структуру скрытого пространства (рисунок 2.2).

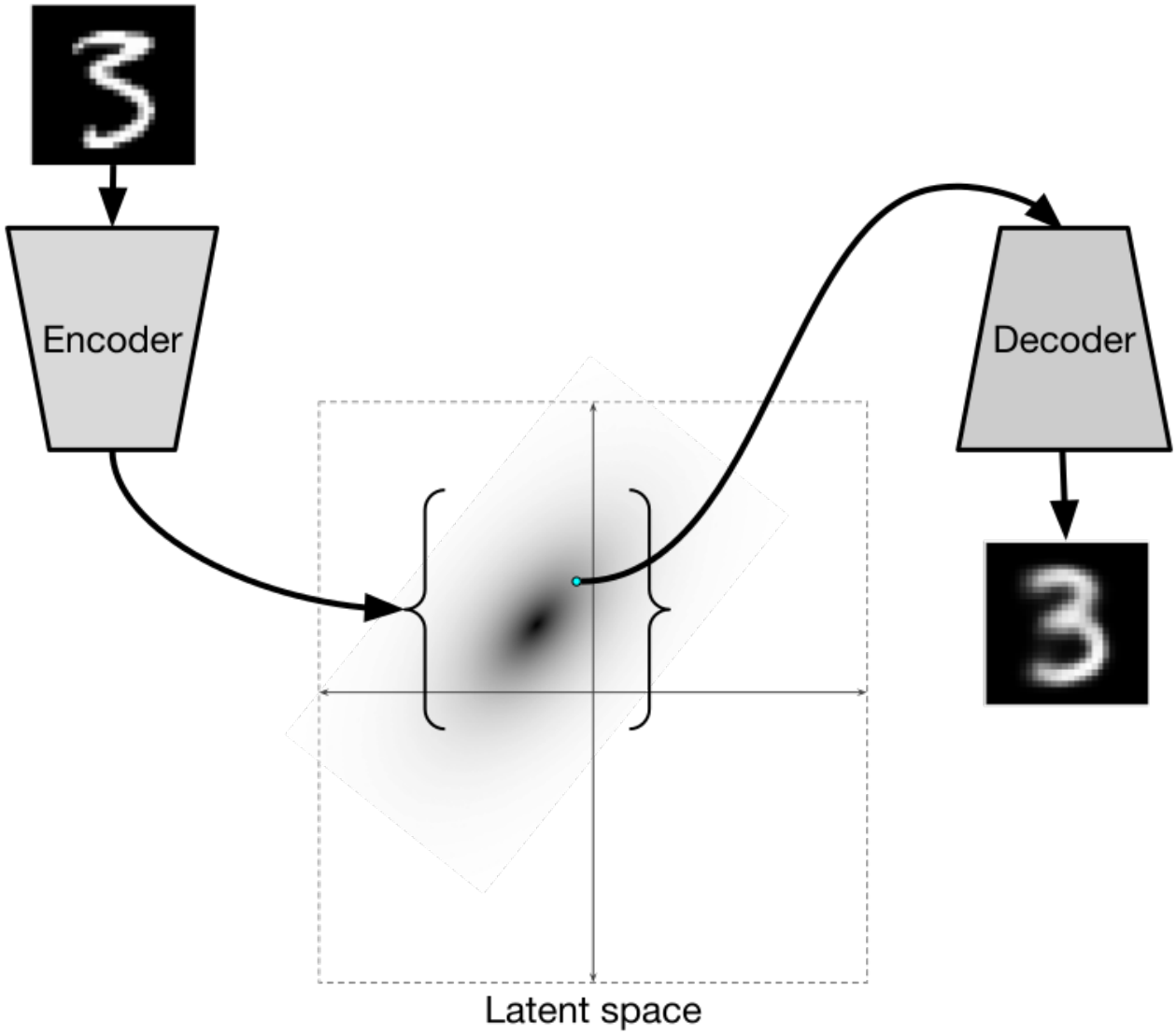

Рисунок 2.2 — Результаты генерации с помощью VAE - видна характерная размытость

Попытки решения проблемы размытости в VAE включали различные модификации архитектуры и функций потерь. WAE (Wasserstein Auto-Encoders) заменяют KL-дивергенцию на расстояние Вассерштейна. VQ-VAE (Vector-Quantized VAE) дискретизируют скрытое пространство для избежания проблем с непрерывными распределениями. Однако все эти модификации лишь частично решают фундаментальную проблему размытости.

Ограничения VAE становятся особенно критичными в задачах, требующих высокой точности деталей, таких как генерация карт по спутниковым снимкам. В картографических приложениях размытость может привести к потере важной географической информации, неточности в границах объектов и общему снижению качества картографического продукта.

Генеративно-состязательные сети представляют принципиально иной под-

ход к генерации данных, который позволяет преодолеть многие ограничения VAE. Основные преимущества GAN перед VAE включают:

Отсутствие явной функции реконструкции. GAN не пытаются точно воспроизвести входные данные, вместо этого они учатся генерировать данные, неотличимые от реальных. Это избавляет от проблемы усреднения, присущей MSE-потере в VAE.

Состязательное обучение обеспечивает резкость. Дискриминатор в GAN принуждает генератор создавать четкие и реалистичные изображения, поскольку размытые изображения легко отличить от реальных. Это создает естественное давление в сторону генерации высококачественных деталей.

Прямая генерация без промежуточного представления. В отличие от VAE, которые должны сначала изучить хорошее скрытое представление, GAN могут сразу фокусироваться на качестве финального изображения.

Возможность использования перцептуальных потерь. GAN позволяют естественно интегрировать более сложные функции потерь, учитывающие восприятие человека, что критично для визуальных приложений.

Лучшая производительность для условной генерации. Условные GAN (cGAN) показывают превосходные результаты в задачах image-to-image translation, включая генерацию карт по спутниковым снимкам.

Таким образом, хотя VAE заложили важные теоретические основы генеративного моделирования и остаются полезными для некоторых приложений (особенно тех, где важно изучение скрытых представлений), для задач генерации высококачественных изображений, таких как создание карт по спутниковым снимкам, GAN демонстрируют значительно лучшие результаты. Переход от VAE к GAN ознаменовал новую эру в генеративном моделировании, особенно в областях, требующих высокой визуальной точности и детализации.

2.2 Теория и архитектура генеративно-состязательных сетей

Генеративно-состязательные сети (Generative Adversarial Networks, GAN) представляют собой мощный класс нейронных сетей, предложенный Иэном Гудфеллоу и коллегами в 2014 году. Идея GAN возникла в результате стремления создать эффективный метод генерации данных, который мог бы обучаться на неразмеченных наборах данных и генерировать новые, реалистич-

ные примеры без проблем размытости, характерных для VAE.

Основная идея GAN заключается в одновременном обучении двух нейронных сетей, которые состязаются друг с другом в рамках игры с нулевой суммой. Стандартная архитектура GAN состоит из генератора $G(z;\theta_g)$, который преобразует случайный шум z из некоторого распределения $p_z(z)$ в данные x, и дискриминатора $D(x;\theta_d)$, который пытается отличить реальные данные от сгенерированных.

Целевая функция GAN формулируется как минимакс игра:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \quad (2.2)$$

Здесь $p_{data}(x)$ представляет истинное распределение данных, $p_z(z)$ - априорное распределение шума (обычно нормальное), $D(x)$ - вероятность того, что x является реальными данными, а $G(z)$ - сгенерированные данные (рисунок 2.3).

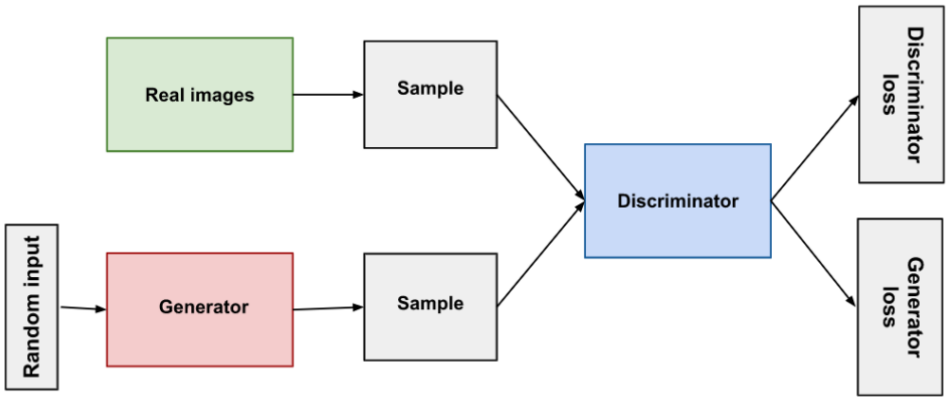

Рисунок 2.3 — Архитектура генеративно-состязательных сетей

Архитектура генератора в GAN разнообразна в зависимости от конкретной задачи и данных. На вход генератора поступает случайный шум или другие формы входных данных. Первый слой генератора может быть плотным слоем (fully connected), который принимает случайный шум и преобразует его в предварительное представление. Слой батч-нормализации используется для стабилизации и ускорения обучения, нормализуя активации между слоями. Слой транспонированной свертки используется для увеличения пространственного разрешения данных, позволяя генератору создавать более сложные структуры, например, изображения большего размера. Выходной слой генератора генерирует финальные сгенерированные данные, которые обычно имеют тот же формат, что и реальные данные.

Архитектура дискриминатора также имеет свои характерные особенности. На вход дискриминатора подаются данные для оценки, которые могут быть как реальными данными из обучающего набора, так и сгенерированными данными от генератора. Дискриминатор обычно начинается с одного или нескольких слоев свертки, которые извлекают признаки из входных данных. После каждого слоя свертки могут следовать функции активации, такие как Leaky ReLU, и слои батч-нормализации для внесения нелинейности и стабилизации обучения. Операция пулинга или субдискретизации используется для уменьшения пространственного разрешения данных. Плоский слой преобразует данные в одномерный вектор для передачи в полносвязные слои. Окончательные слои дискриминатора - полносвязные, которые принимают одномерный вектор и выдают финальное предсказание. Выходной слой дискриминатора часто содержит один нейрон с сигмоидальной функцией активации, вырабатывающий вероятность того, что входные данные являются реальными.

Процесс обучения GAN происходит путем попеременной оптимизации генератора и дискриминатора. В фазе генерации генератор создает данные и передает их дискриминатору, который оценивает их и предоставляет вероятность того, что они реальные. В фазе оценки дискриминатор получает как реальные данные из обучающего набора, так и сгенерированные данные от генератора, обучаясь отличать между ними и обновляя свои веса. В фазе обновления генератора генератор получает обратную связь от дискриминатора и обновляет свои веса таким образом, чтобы создавать более реалистичные данные. Эти фазы повторяются многократно до тех пор, пока генератор не станет способен создавать данные, неотличимые от реальных (рисунок 2.4).

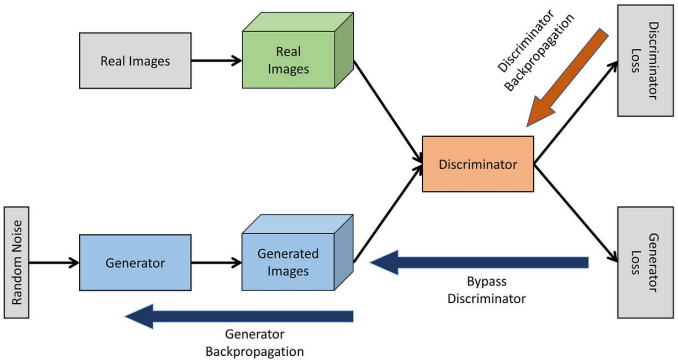

Рисунок 2.4 — Процесс обучения генеративно-состязательных сетей

Однако обучение GAN может сопровождаться рядом особенностей и проблем, которые важно понимать для успешной реализации. Нестабильность обучения является одной из главных проблем GAN. Обучение может быть подвержено нестабильности, что проявляется в изменчивости качества генерации и трудностях в достижении стабильной сходимости между генератором и дискриминатором. Это происходит из-за несбалансированного обучения компонентов: если одна из сторон обучается быстрее или более эффективно, это может вызвать дисбаланс и затруднения в достижении сходимости.

Проблема исчезающего градиента особенно актуальна для генеративно-состязательных сетей. Это явление, при котором градиенты, передаваемые от функции потерь к параметрам модели, становятся крайне маленькими, что затрудняет или делает практически невозможным обновление весов сети в процессе обучения. Градиенты используются для обучения как генератора, так и дискриминатора, и нестабильность обучения может усиливать эту проблему.

Mode Collapse или схлопывание режимов представляет серьезную проблему для GAN. Это явление, при котором генератор создает ограниченное количество уникальных примеров, игнорируя разнообразие в данных и ограничиваясь генерацией ограниченного набора шаблонов или объектов. В результате GAN может потерять способность генерировать разнообразные и реалистичные примеры. Причиной Mode Collapse может послужить нарушение баланса между генератором и дискриминатором: генератор фокусируется на создании примеров, которые «обманывают» дискриминатор, но при этом не обязательно разнообразны.

В процессе обучения GAN могут возникнуть выбросы, что означает создание генератором экстремально необычных примеров, которые сильно отличаются от данных в обучающем наборе. Выбросы могут сказаться на качестве генерации, вводя в данные нехарактерные элементы и снижая их реалистичность.

Существует множество различных типов генеративно-состязательных нейронных сетей, каждый из которых разработан для решения определенных задач и применяется в различных областях. Стандартные GAN (vanilla GAN) представляют оригинальную архитектуру, включающую генератор и дискриминатор, и используются для генерации изображений, звуков, текста и других типов данных. DCGAN (Deep Convolutional GAN) использует сверточные слои в генераторе и дискриминаторе для обработки изображений и применя-

ется для генерации высококачественных изображений в компьютерном зрении.

Особое место занимают условные GAN (Conditional GAN или cGAN), которые представляют расширение стандартного GAN с добавлением условия (метки), определяющего класс или характеристику генерируемых данных. CycleGAN состоит из пары генераторов и дискриминаторов, позволяющих преобразование изображений из одной области в другую и обратно. BigGAN представляет масштабируемую архитектуру GAN с большим количеством параметров, применяемую для генерации высокоразрешенных изображений. Pix2Pix (Image-to-Image GAN) специально предназначен для преобразования изображений из одного пространства в другое и применяется в задачах перевода стилей.

Условные GAN расширяют стандартную архитектуру GAN, включая дополнительную информацию y как для генератора, так и для дискриминатора. Эта дополнительная информация может представлять собой метки классов, текстовые описания или, в случае pix2pix, входные изображения. Целевая функция условных GAN модифицируется следующим образом:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)] + \\ \mathbb{E}_{z\sim p_z(z),y\sim p_{data}(y)}[\log(1-D(G(z,y),y))],\end{aligned} \tag{2.3}$$

Где y представляет условную информацию. В случае pix2pix, y соответствует входному изображению, а x - целевому изображению.

Условные GAN обладают рядом принципиальных преимуществ для задач image-to-image translation. Они обеспечивают контролируемую генерацию, позволяя направлять процесс генерации через входные условия. Кроме того, они способны сохранять структуру, поддерживая пространственное соответствие между входом и выходом, и обеспечивают семантическую согласованность между условием и результатом. Это особенно важно для картографических приложений, где необходимо сохранить точное пространственное соответствие между спутниковым снимком и генерируемой картой.

2.3 Архитектура и реализация модели pix2pix

Модель pix2pix, предложенная Изола и коллегами в 2017 году, представляет собой специализированную архитектуру условных GAN, разработанную специально для задач image-to-image translation. Ключевые особенно-

сти pix2pix включают использование архитектуры U-Net для генератора и PatchGAN для дискриминатора, что обеспечивает эффективное преобразование спутниковых снимков в картографические изображения.

Функция потерь pix2pix сочетает состязательную потерю и L1 регуляризацию:

$$\mathcal{L}_{pix2pix}(G, D) = \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G), \tag{2.4}$$

Здесь $\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]$ представляет условную состязательную потерю, а $\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x, z)||_1]$ - L1 потерю. Гиперпараметр λ (обычно равный 100) балансирует вклад L1 потери.

L1 регуляризация играет критическую роль в обеспечении низкочастотной корректности, заставляя генератор создавать изображения, близкие к целевым на пиксельном уровне. Это особенно важно для картографических приложений, где точность пространственного позиционирования объектов имеет первостепенное значение. Комбинация состязательной потери с L1 регуляризацией позволяет сочетать реалистичность генерируемых изображений с их точностью на пиксельном уровне.

Архитектура U-Net генератора представляет собой одну из наиболее элегантных и эффективных конструкций для задач, требующих сохранения пространственной информации. U-Net была первоначально разработана Ронневергером и коллегами в 2015 году для задач биомедицинской сегментации и получила свое название благодаря характерной U-образной форме архитектуры.

Основная идея U-Net заключается в использовании симметричной энкодер-декодерной архитектуры с пропускными соединениями. Энкодер (левая часть "U") последовательно уменьшает пространственное разрешение, увеличивая количество каналов и глубину представления. Декодер (правая часть "U") восстанавливает пространственное разрешение, используя транспонированные свертки.

Энкодер состоит из последовательности сверточных блоков, каждый из которых выполняет следующие операции. Сначала применяется 2D-свертка с ядром размера 4×4, шагом 2 и паддингом 1, что уменьшает пространственные размеры в два раза и позволяет выделить более абстрактные признаки. Затем используется функция активации Leaky ReLU с коэффициентом отрицательного наклона 0.2, которая позволяет небольшому градиенту проходить

через отрицательные значения, предотвращая проблему "мертвых нейронов". Нормализация батча применяется после первого слоя для стабилизации обучения и ускорения конвергенции.

Математически, каждый слой энкодера может быть описан как:

$$h_{i+1} = \text{LeakyReLU}(\text{BN}(\text{Conv2d}(h_i))), \tag{2.5}$$

где h_i - выход i-го слоя энкодера.

Архитектура энкодера в pix2pix включает восемь последовательных блоков, где количество каналов увеличивается по схеме: $64 \to 128 \to 256 \to 512 \to 512 \to 512 \to 512 \to 512$. Это прогрессивное увеличение количества каналов позволяет модели изучать все более сложные и абстрактные представления входного изображения.

Декодер U-Net выполняет обратные операции, восстанавливая пространственное разрешение изображения. Он использует транспонированные свертки (ConvTranspose2d) с ядром размера 4×4, шагом 2 и паддингом 1 для увеличения разрешения в два раза на каждом шаге. Функции активации ReLU используются для внесения нелинейности, в отличие от Leaky ReLU в энкодере. Нормализация батча применяется для стабилизации обучения. Dropout с вероятностью 0.5 используется в первых трех слоях декодера для предотвращения переобучения.

Количество каналов в декодере уменьшается симметрично энкодеру, но с учетом пропускных соединений: $512 \to 512 \to 512 \to 512 \to 256 \to 128 \to 64 \to \text{output\_channels}$.

Ключевой особенностью U-Net являются пропускные соединения, которые соединяют соответствующие уровни энкодера и декодера:

$$d_i = \text{Concat}(u_i, e_{n-i}), \tag{2.6}$$

где d_i - выход i-го слоя декодера, u_i - апсэмплированный выход предыдущего слоя декодера, и e_{n-i} - соответствующий слой энкодера.

Эти соединения выполняют несколько критических функций. Во-первых, они передают низкоуровневую пространственную информацию напрямую от энкодера к декодеру, минуя узкое место (bottleneck) архитектуры. Во-вторых, они позволяют комбинировать локальные детали с глобальным контекстом, что критично для точной генерации карт. В-третьих, они улучшают градиентный поток во время обучения, уменьшая проблему исчезающих градиентов.

С точки зрения теории информации, пропускные соединения решают фундаментальную проблему информационного bottleneck. В традиционных энкодер-декодерных архитектурах узкое место вынуждает всю пространственную информацию проходить через ограниченное представление, что неизбежно приводит к потере деталей. Пропускные соединения создают альтернативные пути для передачи информации, позволяя сохранить детали различного масштаба.

Архитектура дискриминатора PatchGAN представляет инновационный подход к оценке качества изображений. Вместо классификации всего изображения как реального или поддельного, PatchGAN принимает решения на уровне локальных патчей размером $N \times N$. Этот подход основан на наблюдении, что реалистичность изображения может быть эффективно оценена на локальном уровне, а глобальная структурная корректность обеспечивается L1 потерей.

PatchGAN можно математически описать как:

$$D(x, y) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} D_p(x, y), \tag{2.7}$$

где $\mathcal{P}$ - множество всех патчей в изображении, а $D_p(x, y)$ - решение дискриминатора для конкретного патча p.

Архитектура дискриминатора состоит из серии базовых блоков, каждый из которых включает сверточный слой Conv2d с ядром 4×4, шагом 2 и паддингом 1, нормализацию InstanceNorm (вместо BatchNorm для лучшей работы с малыми батчами), и функцию активации Leaky ReLU с параметром 0.2.

Входной слой дискриминатора выполняет конкатенацию условного изображения (спутниковый снимок) и целевого изображения (реальная или сгенерированная карта) по канальному измерению. Это создает вход размером 6 каналов (3 канала RGB для каждого изображения), что позволяет дискриминатору учитывать соответствие между входным условием и выходным результатом.

Последовательность сверточных блоков прогрессивно уменьшает пространственное разрешение и увеличивает количество каналов: $6 \to 64 \to 128 \to 256 \to 512$. Важно, что первый блок не использует нормализацию, что позволяет сохранить исходные характеристики входных изображений. Финальный сверточный слой с ядром 4×4, шагом 1 и паддингом 1 создает карту активации размером 30×30, где каждый элемент представляет классификацию

соответствующего патча в исходном изображении.

PatchGAN обладает рядом архитектурных преимуществ. Достаточно малое количество параметров по сравнению с полноразмерным дискриминатором делает обучение более эффективным и требует меньше памяти. Фокус на локальных текстурах и деталях критически важен для реалистичности изображений, поскольку многие артефакты генерации проявляются именно на локальном уровне. Инвариантность к размеру входного изображения позволяет использовать одну модель для данных различного разрешения. Локальная оценка качества особенно важна для картографических изображений, где точность воспроизведения каждого региона критична для правильной интерпретации географической информации.

Процесс обучения pix2pix требует тщательной балансировки между генератором и дискриминатором. Обучение происходит по итеративному алгоритму, где на каждой итерации сначала обновляется дискриминатор при фиксированном генераторе, затем обновляется генератор при фиксированном дискриминаторе.

Фаза обучения дискриминатора включает в себя вычисление потери $D_{loss} = \mathcal{L}_{cGAN}(D, G)$ на реальных и сгенерированных парах изображений. Дискриминатор стремится максимизировать вероятность правильной классификации реальных изображений как настоящих, а сгенерированных как поддельных.

Фаза обучения генератора минимизирует комбинированную потерю $G_{loss} = \mathcal{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{L1}(G)$. Генератор стремится обмануть дискриминатор, создавая изображения, которые дискриминатор классифицирует как реальные, одновременно минимизируя L1 расстояние до целевого изображения.

Ключевые параметры обучения тщательно подобраны для обеспечения стабильной конвергенции. Оптимизатор Adam используется с параметрами $\beta_1 = 0.5$ и $\beta_2 = 0.999$, где уменьшенное значение β_1 помогает стабилизировать обучение GAN за счет уменьшения инерции момента. Скорость обучения 2×10^{-4} установлена одинаковой для обеих сетей, что обеспечивает сбалансированное обучение. Размер батча ограничен значениями 1-4 из-за ограничений GPU памяти при работе с изображениями высокого разрешения.

Инициализация весов выполняется из нормального распределения с нулевым средним и стандартным отклонением $\sigma = 0.02$. Эта инициализация оказалась эмпирически эффективной для GAN и помогает стабилизировать

начальную фазу обучения.

Стабилизация процесса обучения достигается через несколько техник. Чередующееся обучение компонентов помогает поддерживать динамическое равновесие между генератором и дискриминатором. Возможность использования разных скоростей обучения позволяет тонко настраивать этот баланс в зависимости от конкретной задачи. Постепенное снижение скорости обучения (learning rate scheduling) позволяет модели точнее настроиться в поздних стадиях обучения. Использование экспоненциально скользящего среднего (EMA) для весов генератора может помочь получить более стабильные результаты.

Критически важным аспектом является баланс между состязательной потерей и L1 потерей. Слишком большой вес L1 потери ($\lambda > 1000$) может привести к генерации размытых изображений, поскольку L1 потеря стремится к среднему значению всех возможных выходов. Недостаточный вес L1 потери ($\lambda < 10$) может привести к реалистичным, но пространственно несогласованным результатам. Значение $\lambda = 100$ было определено эмпирически как оптимальное для большинства задач image-to-image translation, включая генерацию карт.

Мониторинг процесса обучения включает отслеживание нескольких ключевых метрик. Потери генератора и дискриминатора должны снижаться и стабилизироваться, но не обязательно стремиться к нулю. Качественная оценка сгенерированных изображений на валидационном наборе помогает выявить проблемы, такие как mode collapse или артефакты генерации. Регулярная оценка количественных метрик (SSIM, PSNR, FID) позволяет объективно отслеживать прогресс обучения.

Важным аспектом является время обучения и критерии остановки. Обычно модель достигает приемлемого качества после 100-200 эпох, но окончательное решение должно основываться на стабильности метрик качества и отсутствии признаков переобучения или недообучения. Ранняя остановка может применяться при отсутствии улучшения на валидационном наборе в течение определенного количества эпох.

2.4 Результаты обучения и анализ производительности

Обучение модели pix2pix демонстрирует стабильную конвергенцию, что подтверждается графиками функций потерь (рисунок 2.5). Наблюдается по-

степенное снижение потерь генератора и дискриминатора с достижением равновесия после примерно 100-200 эпох обучения. Характерно, что потери не стремятся к нулю, а стабилизируются на определенном уровне, что указывает на достижение динамического равновесия между генератором и дискриминатором. Успешная конвергенция проявляется в том, что дискриминатор перестает легко отличать сгенерированные изображения от реальных, в то время как генератор продолжает создавать качественные результаты.

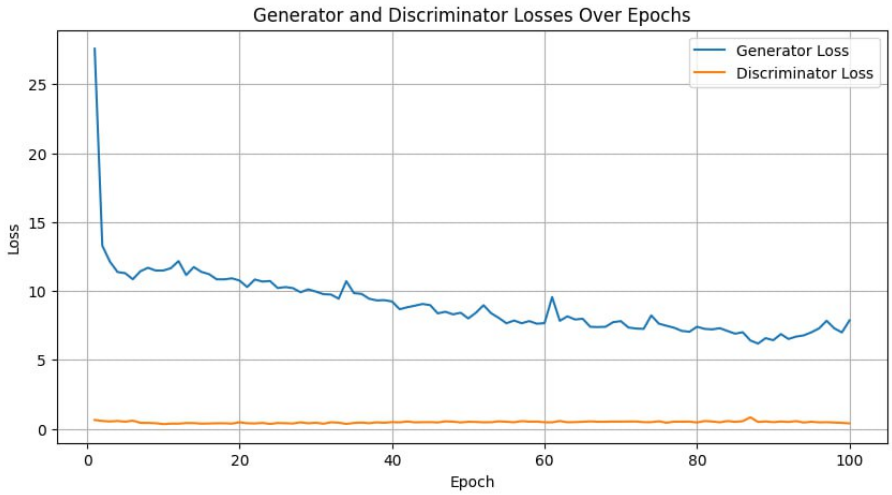

Рисунок 2.5 — Динамика функций потерь генератора и дискриминатора в процессе обучения

Визуальное сравнение результатов демонстрирует способность модели генерировать карты, сохраняющие основные структурные элементы входных спутниковых изображений (рисунок 2.6). Модель показывает особенно хорошие результаты в воспроизведении четко определенных объектов: дороги сохраняют свою связность и правильную топологию, здания воспроизводятся с сохранением их формы и размеров, водные объекты генерируются с четкими границами, растительность передается с соответствующими цветами и текстурами. Однако наблюдается некоторое упрощение мелких деталей и иногда размытие в областях со сложной структурой.

Предварительная количественная оценка показывает обнадеживающие результаты. SSIM находится в диапазоне 0.75-0.85, что указывает на хорошее структурное сходство между сгенерированными и реальными картами с сохранением основных визуальных паттернов. PSNR достигает уровня 25-30 dB, демонстрируя приемлемое соотношение сигнал-шум и качество восстановления изображения. L1 Loss находится в пределах 0.15-0.25, что свидетельствует о относительно низкой средней абсолютной ошибке между сгенерированными и целевыми изображениями на пиксельном уровне.

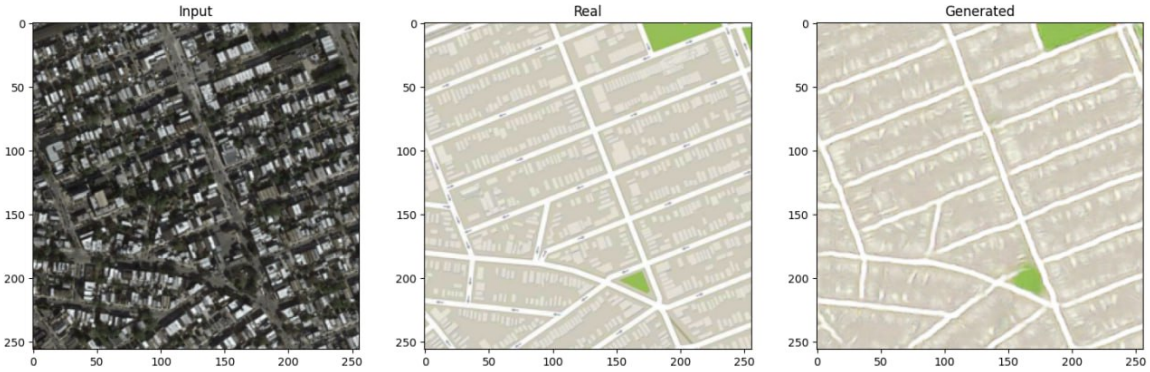

Рисунок 2.6 — Сравнение исходных спутниковых снимков, сгенерированных карт и реальных карт

Несмотря на успешные результаты, модель pix2pix имеет несколько заметных ограничений, которые важно учитывать при практическом применении. Наблюдается характерная тенденция к генерации слегка размытых изображений, особенно в областях с мелкими деталями, такими как мелкие дороги или границы небольших объектов. Сверточная природа архитектуры ограничивает глобальное понимание контекста, что может приводить к локальным несогласованностям в крупномасштабных структурах, таких как речные системы или дорожные сети большой протяженности.

Иногда появляются нереалистичные элементы или артефакты генерации, особенно на границах между различными типами объектов, где модель может создавать переходные зоны, не соответствующие реальности. Модель требует высококачественных парных данных для эффективного обучения и может плохо генерализоваться на данные, значительно отличающиеся от обучающих примеров по условиям съемки, временным характеристикам или географическому расположению.

2.5 Выводы к главе 2

Модель pix2pix представляет собой эффективное и проверенное временем решение для задач image-to-image translation, включая генерацию карт по спутниковым снимкам. Комбинация архитектуры U-Net для генератора и PatchGAN для дискриминатора обеспечивает хороший баланс между сохранением детальной пространственной информации и созданием реалистичных результатов.

Ключевые достоинства подхода многообразны и хорошо документированы. Модель демонстрирует стабильность обучения и надежную конверген-

цию, что делает её привлекательной для практического применения. Хорошее сохранение пространственной структуры входных данных обеспечивается благодаря пропускным соединениям U-Net и использованию L1 регуляризации. Относительная простота реализации и воспроизводимость результатов делают pix2pix доступным для широкого круга исследователей и практиков. Разумная вычислительная эффективность при умеренных требованиях к ресурсам позволяет использовать модель даже на ограниченном оборудовании.

Модель демонстрирует особенно хорошие результаты в воспроизведении четко определенных структур, таких как дороги, здания и водные объекты, что критически важно для картографических приложений. Использование PatchGAN дискриминатора обеспечивает хорошее качество локальных деталей и текстур, что повышает визуальную реалистичность генерируемых карт.

Однако ограничения сверточных архитектур в моделировании дальних зависимостей и глобального контекста открывают перспективы для исследования более продвинутых подходов. Тенденция к размытию мелких деталей и ограниченное понимание глобальной структуры изображения указывают на необходимость рассмотрения альтернативных архитектур. Эти ограничения становятся особенно заметными при работе с крупномасштабными спутниковыми изображениями, содержащими сложные пространственные структуры и взаимосвязи между удаленными объектами.

В следующей главе будет детально рассмотрена экспериментальная архитектура, интегрирующая Vision Transformer в генеративные модели для потенциального преодоления этих ограничений и достижения более высокого качества генерации карт. Особое внимание будет уделено способности трансформеров моделировать дальние зависимости и глобальный контекст, что может привести к значительному улучшению качества генерации, особенно для объектов большого протяжения и сложных пространственных структур.

ГЛАВА 3

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ И РЕАЛИЗАЦИЯ ЭКСПЕРИМЕНТАЛЬНОЙ МОДЕЛИ GAN С ГЕНЕРАТОРОМ НА VISION TRANSFORMER

3.1 Теоретические основы Vision Transformer

Vision Transformer (ViT) представляет собой революционную адаптацию архитектуры Transformer, изначально разработанной для задач обработки естественного языка, к задачам компьютерного зрения. Появление ViT ознаменовало принципиальный сдвиг в подходах к анализу изображений, предложив альтернативу доминировавшим до этого сверточным нейронным сетям.

Фундаментальная идея ViT заключается в том, что изображение можно рассматривать как последовательность токенов - патчей фиксированного размера, которые могут быть обработаны механизмом внимания аналогично словам в тексте. Такой подход позволяет применить мощные возможности трансформеров для моделирования дальних зависимостей к визуальному контенту.

Входное изображение размером $H \times W \times C$ разбивается на $N = \frac{H \times W}{P^2}$ неперекрывающихся патчей размером $P \times P$, где P обычно выбирается равным 16. Каждый патч рассматривается как отдельный токен и проецируется в векторное пространство размерности D с помощью линейного слоя (обычно реализованного как сверточный слой с размером ядра $P \times P$ и шагом P).

Критически важным компонентом ViT является позиционное кодирование, которое добавляется к патч-эмбеддингам для сохранения информации о пространственном расположении патчей. В отличие от текстовых данных, где позиции слов имеют четко определенный порядок, изображения имеют двумерную структуру, что требует специальных схем кодирования пространственных отношений. Последовательность патч-эмбеддингов с позиционным кодированием формируется следующим образом:

$$z_0 = [x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}, \tag{3.1}$$

где $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ - матрица проекции патчей, а $E_{pos} \in \mathbb{R}^{N \times D}$ - позиционные эмбеддинги.

Ядром архитектуры ViT является механизм многоголового самовнимания (Multi-Head Self-Attention), который позволяет каждому патчу взаимодействовать со всеми остальными патчами изображения независимо от их пространственного расположения. Математически механизм внимания определяется как:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.2}$$

где $Q = zW_Q$, $K = zW_K$, $V = zW_V$ - проекции входных токенов в пространства запросов, ключей и значений соответственно. Размерность d_k используется для нормализации скалярных произведений, предотвращая насыщение функции softmax.

Многоголовое внимание расширяет эту концепцию, используя h параллельных голов внимания, каждая из которых работает в своем подпространстве:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{3.3}$$

где $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ и W^O - выходная проекция для объединения результатов всех голов.

Трансформерный блок состоит из слоя многоголового самовнимания, за которым следует position-wise feed-forward network (MLP). Оба компонента окружены residual соединениями и layer normalization:

$$\begin{aligned} z_l'' &= z_l + \text{MultiHead}(\text{LayerNorm}(z_l)), \\ z_{l+1} &= z_l'' + \text{MLP}(\text{LayerNorm}(z_l'')), \end{aligned} \tag{3.4}$$

MLP-блок обычно состоит из двух линейных преобразований с нелинейной активационной функцией GELU между ними:

$$\text{MLP}(x) = \text{Linear}(\text{GELU}(\text{Linear}(x))), \tag{3.5}$$

Принципиальное преимущество ViT перед сверточными сетями заключается в способности моделировать глобальные зависимости с помощью механизма внимания. В то время как сверточные сети ограничены локальными

рецептивными полями и должны использовать множество слоев для агрегации информации с больших областей изображения, ViT может напрямую устанавливать связи между любыми двумя патчами за один шаг.

Для задач генерации карт по спутниковым снимкам эта особенность ViT представляется особенно ценной. Спутниковые изображения часто содержат протяженные структуры, такие как дороги, реки, границы ландшафтов, которые простираются через всё изображение. Традиционные сверточные сети могут испытывать трудности в сохранении связности таких структур из-за ограниченности локальных рецептивных полей. ViT, благодаря механизму внимания, способен одновременно учитывать весь контекст изображения и более эффективно моделировать такие глобальные структуры.

3.2 Интеграция Vision Transformer в архитектуру GAN

Интеграция Vision Transformer в генеративно-состязательные сети для задач image-to-image translation представляет собой нетривиальную техническую задачу, требующую решения фундаментальной проблемы адаптации дискретной патч-ориентированной обработки ViT к непрерывной генерации изображений высокого разрешения.

Ключевая проблема заключается в том, что классическая архитектура ViT была разработана для задач классификации, где требуется получить единственное предсказание для всего изображения. В задачах генерации необходимо создать выходное изображение того же разрешения, что и входное, сохраняя при этом детальную пространственную структуру. Это требует принципиально иного подхода к организации вычислительного потока в трансформерной архитектуре.

Предлагаемое решение основано на гибридной архитектуре, которая сочетает способность ViT к глобальному контекстному анализу с эффективностью сверточных операций для восстановления пространственного разрешения. Генератор реализует отображение $G : X \to Y$ через специализированную последовательность трансформерных и сверточных операций.

В отличие от классификационных архитектур, где используется специальный classification token, все патчи обрабатываются равноправно для сохранения пространственной информации. Критическим аспектом является сохранение пространственной структуры токенов на всех этапах обработки.

Переходный блок между трансформерной частью и CNN-декодером пред-

ставляет собой инновационное решение проблемы преобразования дискретных токенов в непрерывные карты признаков. Этот блок включает линейную проекцию для изменения размерности и специальную операцию rearrange для восстановления двумерной структуры.

Ключевой инновацией является специализированная архитектура CNN-декодера, разработанная для эффективной обработки признаков, извлеченных трансформером. Декодер использует прогрессивное увеличение разрешения с одновременным уточнением локальных деталей, что позволяет максимально использовать глобальную информацию, полученную от трансформера.

Особенностью интеграции является отсутствие прямых skip connections между трансформерной частью и декодером, в отличие от U-Net архитектуры. Это решение обусловлено различием в природе обработки: трансформер оперирует дискретными токенами с глобальными взаимосвязями, в то время как декодер работает с локальными иерархическими представлениями.

Дискриминатор сохраняет архитектуру PatchGAN без модификаций, поскольку его функциональность не зависит от типа генератора. Функция потерь также остается идентичной pix2pix, что обеспечивает справедливое сравнение архитектур.

3.3 Архитектура генератора на основе Vision Transformer

Детальная архитектура генератора представляет собой тщательно спроектированную систему, оптимизированную для эффективной обработки пространственной информации спутниковых изображений. Каждый компонент архитектуры имеет специфические особенности, адаптированные для задач генерации карт.

Входной блок патч-эмбеддинга реализован с помощью сверточного слоя с ядром 16×16 и соответствующим шагом, что обеспечивает оптимальное разбиение изображения 256×256 на 256 токенов. Проекция в пространство размерности 512 была выбрана как компромисс между выразительной способностью и вычислительной эффективностью.

Позиционное кодирование использует схему learnable эмбеддингов, обучаемых совместно с остальными параметрами модели. Эта схема показала лучшую адаптацию к специфике двумерных изображений по сравнению с синусоидальным кодированием, заимствованным из NLP.

Трансформерная часть включает 6 последовательных блоков с 8 головами внимания каждый. Pre-normalization архитектура обеспечивает более стабильное обучение для генеративных задач. MLP-блоки имеют расширенную размерность (2048), что обеспечивает достаточную нелинейность для сложных преобразований.

Анализ весов внимания выявляет интересные паттерны: некоторые головы специализируются на локальных взаимодействиях между соседними патчами, другие устанавливают дальние связи между семантически связанными областями. Эта специализация возникает естественно в процессе обучения без явного руководства.

Переходный блок решает нетривиальную задачу преобразования последовательности токенов в двумерную карту признаков. Операция rearrange восстанавливает пространственную структуру согласно формуле '(h w) d -> h w d', где h и w соответствуют количеству патчей по высоте и ширине.

CNN-декодер спроектирован для четырехэтапного увеличения разрешения от $16 \times 16 \times 512$ до $256 \times 256 \times 3$. Каждый этап включает транспонированную свертку для upsampling и серию residual блоков для уточнения признаков. Использование residual connections критически важно для предотвращения исчезающих градиентов в декодере.

Архитектурной особенностью является отсутствие skip connections между трансформерной частью и декодером. Попытки их введения приводили к нестабильности обучения из-за несовместимости дискретной природы токенов и непрерывных карт признаков CNN.

Регуляризация реализуется через dropout в MLP-блоках трансформера (вероятность 0.1) и batch normalization в декодере. Финальный слой использует активацию tanh для ограничения выходных значений в диапазоне [-1, 1], соответствующем нормализации данных.

3.4 Особенности реализации и процесса обучения

Практическая реализация ViT-GAN требует специализированных подходов к обучению, существенно отличающихся от стандартных практик для сверточных архитектур. Трансформеры демонстрируют уникальные характеристики в процессе обучения, требующие тщательного внимания к стабилизации и мониторингу.

Выбор оптимизатора AdamW с weight decay обусловлен склонностью транс-

формеров к переобучению весов внимания. В отличие от стандартного Adam, AdamW обеспечивает более эффективную регуляризацию за счет разделения обновления весов и weight decay. Параметры $\beta_1 = 0.5$ адаптированы специально для GAN обучения, обеспечивая баланс между стабильностью генератора и дискриминатора.

Косинусоидальное расписание скорости обучения с warm-up периодом критически важно для стабилизации начальных этапов обучения. Механизм внимания особенно чувствителен к инициализации, и резкие изменения в начале обучения могут привести к коллапсу или расходимости:

$$lr_{warmup}(t) = lr_{target} \cdot \min\left(1, \frac{t}{T_{warmup}}\right), \tag{3.6}$$

Gradient accumulation имитирует большие размеры батчей при ограниченной GPU памяти. Трансформеры особенно выигрывают от больших батчей из-за нормализации в механизме внимания. Эффективный размер батча 16 (4 $\times$ accumulation-steps) показал оптимальный баланс между стабильностью и вычислительными требованиями.

Ограничение нормы градиентов предотвращает gradient explosion, характерный для глубоких трансформерных сетей. Максимальная норма 1.0 была определена эмпирически как компромисс между предотвращением резких скачков и сохранением информативности градиентов.

Смешанная точность с автоматическим масштабированием градиентов снижает потребление памяти на 40-50% без заметной потери качества. Трансформеры хорошо адаптированы к float16 вычислениям благодаря нормализации слоев и residual connections.

Критической особенностью является повышенная нестабильность обучения ViT-GAN по сравнению с CNN-архитектурами. Механизм внимания может приводить к резким изменениям в паттернах активации, особенно в первые 20-30 эпох. Наблюдается характерный фазовый переход - резкое улучшение качества после начального периода нестабильности.

Мониторинг весов внимания играет ключевую роль в диагностике процесса обучения. Энтропия распределения внимания служит индикатором здоровья модели: слишком низкая энтропия указывает на коллапс внимания, слишком высокая - на недообучение. Оптимальная энтропия варьируется в диапазоне 3.5-4.2 для данной архитектуры.

Управление памятью представляет значительный вызов из-за квадратич-

ной сложности внимания. Gradient checkpointing позволяет обменивать память на вычисления, пересчитывая промежуточные значения во время backward pass. Это увеличивает время обучения на 25-40%, но позволяет работать с ограниченной GPU памятью.

Техника attention dropout с вероятностью 0.1 применяется только к весам внимания для предотвращения переобучения специфическим паттернам. Стандартный dropout в MLP-блоках может дестабилизировать механизм внимания и применяется с осторожностью.

Специфическая проблема ViT-GAN - возможность attention collapse, когда все головы внимания конвергируют к похожим паттернам. Для предотвращения используется диверсификационная регуляризация, штрафующая избыточное сходство между головами внимания.

Валидационный процесс включает не только стандартные метрики, но и анализ карт внимания для понимания фокуса модели. Визуализация показывает, какие области спутникового изображения влияют на генерацию каждого участка карты, что помогает в интерпретации результатов.

Критерии остановки обучения основаны на стабилизации как функций потерь, так и паттернов внимания. Модель считается сходящейся, когда энтропия весов внимания стабилизируется, а метрики качества не улучшаются в течение 15 последовательных эпох на валидационном наборе.

3.5 Выводы к главе 3

Интеграция Vision Transformer в архитектуру генеративно-состязательных сетей для задач генерации карт по спутниковым снимкам представляет собой значительный шаг в развитии методов глубокого обучения для геоинформационных приложений. Предложенная архитектура успешно адаптирует принципы трансформерного внимания к задачам пространственной генерации, преодолевая фундаментальные ограничения традиционных сверточных подходов.

Теоретическое обоснование применения ViT в данной задаче базируется на способности механизма внимания моделировать дальние пространственные зависимости, что критически важно для корректной генерации протяженных географических объектов. В отличие от сверточных сетей, которые агрегируют информацию последовательно через множество слоев, ViT может непосредственно устанавливать связи между любыми областями изображе-

ния, что теоретически должно приводить к лучшему пониманию глобальной структуры спутниковых снимков.

Гибридная архитектура, сочетающая трансформерную обработку признаков с CNN-декодером, представляет собой прагматичное решение проблемы адаптации дискретных токенов к непрерывной генерации изображений. Этот подход позволяет объединить преимущества трансформеров в моделировании глобального контекста с эффективностью сверточных сетей в генерации локальных деталей и текстур.

Особенности реализации и обучения ViT-GAN выявили ряд практических вызовов, включая повышенную чувствительность к гиперпараметрам, большие требования к вычислительным ресурсам и необходимость специализированных техник стабилизации обучения. Эти ограничения отражают фундаментальную сложность трансформерных архитектур и требуют тщательного баланса между теоретическими преимуществами и практической применимостью.

Разработанная методология интеграции ViT в GAN может служить основой для дальнейших исследований в области применения трансформеров к задачам генерации изображений. Особый интерес представляют потенциальные улучшения в архитектуре, такие как адаптивное разбиение на патчи, более эффективные схемы позиционного кодирования и оптимизации механизма внимания для двумерных данных.

Несмотря на технические сложности, предложенный подход открывает новые возможности для создания более точных и детализированных картографических материалов. Способность ViT к глобальному контекстному анализу особенно ценна для обработки сложных спутниковых изображений с множественными пересекающимися объектами и структурами различного масштаба.

В следующей главе будет проведено детальное сравнение разработанной ViT-GAN архитектуры с классическим подходом pix2pix, включающее количественную оценку качества генерации, анализ вычислительной эффективности и практические рекомендации по выбору наиболее подходящего метода для различных сценариев применения.

ГЛАВА 4

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ И ПОДХОДОВ

4.1 Методология экспериментального исследования

Для проведения объективного сравнительного анализа двух архитектур была разработана комплексная методология, обеспечивающая справедливость и воспроизводимость результатов. В качестве экспериментальной базы использовался датасет Maps, содержащий пары изображений спутниковый снимок - карта, который был разделен на обучающую (80%), валидационную (10%) и тестовую (10%) выборки согласно стандартным практикам машинного обучения.

Унифицированный пайплайн предобработки обеспечивал идентичные условия для обеих моделей. Все изображения масштабировались до размера 256×256 пикселей, что соответствует оптимальным параметрам генеративных архитектур. Нормализация значений пикселей выполнялась согласно формуле $I_{norm} = \frac{I - 0.5}{0.5}$, приводя значения к диапазону $[-1, 1]$, оптимизированному для функции активации tanh на выходе генераторов.

Для повышения робастности моделей применялись техники аугментации данных, включающие горизонтальное и вертикальное отражение, вращение на 180 градусов, центральную обрезку с последующим масштабированием. Идентичные трансформации применялись к обоим изображениям пары для сохранения пространственного соответствия. Детерминированность результатов обеспечивалась фиксацией начальных значений генераторов случайных чисел.

Система метрик качества включала множественные объективные показатели, каждый из которых оценивал специфические аспекты генерации. Индекс структурного сходства (SSIM) оценивал визуальное сходство с учетом особенностей человеческого восприятия. Fréchet Inception Distance (FID) измеряла реалистичность сгенерированных изображений через анализ распределения признаков. Intersection over Union (IoU) оценивала точность воспроизведения отдельных элементов карты для различных классов объектов (дороги, здания, водоемы, растительность).

Анализ вычислительной эффективности основывался на измерении времени обучения на одну эпоху, времени инференса для одного изображения, пикового потребления GPU-памяти и количества обучаемых параметров модели. Дополнительно оценивалась стабильность процесса обучения через дисперсию функции потерь на последних эпохах и скорость сходимости.

Для качественного анализа были разработаны инструменты визуализации, включающие генерацию тепловых карт отличий между целевыми и сгенерированными изображениями. Для трансформерной модели особый интерес представлял анализ матриц внимания, показывающий области изображения, между которыми устанавливаются наиболее сильные связи.

4.2 Количественное сравнение эффективности моделей

Результаты экспериментального сравнения базовой архитектуры pix2pix с U-Net генератором и разработанной трансформерной модели представлены в обобщенном виде в таблице 4.1. Данные демонстрируют небольшое, но последовательное превосходство трансформерной модели по всем метрикам качества при существенных различиях в вычислительных требованиях.

Таблица 4.1 — Сравнение эффективности моделей по ключевым метрикам

Метрика	pix2pix (U-Net)	ViT-GAN	Изменение
SSIM ↑	0.876	0.892	+1.8%
FID ↓	58.7	56.8	-3.2%
IoU ↑	0.753	0.772	+2.5%
Время обучения	1.0×	1.8×	+80%
Потребление памяти	1.0×	2.3×	+130%

Трансформерная модель показала улучшение по метрике SSIM на 1.8%, что указывает на лучшее сохранение структурных характеристик изображения. Наиболее значительное улучшение наблюдалось по метрике IoU (+2.5%), свидетельствующее о более точном воспроизведении boundaries объектов на картах. Снижение FID на 3.2% демонстрирует некоторое повышение реалистичности генерируемых изображений.

Однако достижение этих улучшений потребовало значительных вычислительных затрат. Трансформерная модель потребовала в 1.8 раза больше времени на обучение и в 2.3 раза больше GPU памяти. Эти различия объясняются квадратичной сложностью механизма внимания относительно количества токенов и более глубокой архитектурой.

Важным аспектом сравнения является сложность настройки и оптимизации моделей. Базовая архитектура pix2pix продемонстрировала устойчивые результаты "из коробки"с минимальной настройкой гиперпараметров. В противоположность этому, трансформерная модель потребовала тщательного подбора параметров обучения, стратегий регуляризации, схем позиционного кодирования и инициализации весов.

Анализ динамики обучения выявил различные паттерны конвергенции. Модель pix2pix характеризовалась плавной и монотонной сходимостью функции потерь, в то время как трансформерная модель демонстрировала более резкие колебания, особенно в начальные эпохи (рисунок 4.1). Это объясняется необходимостью стабилизации множественных механизмов внимания, которые одновременно оптимизируют взаимосвязи между всеми элементами изображения.

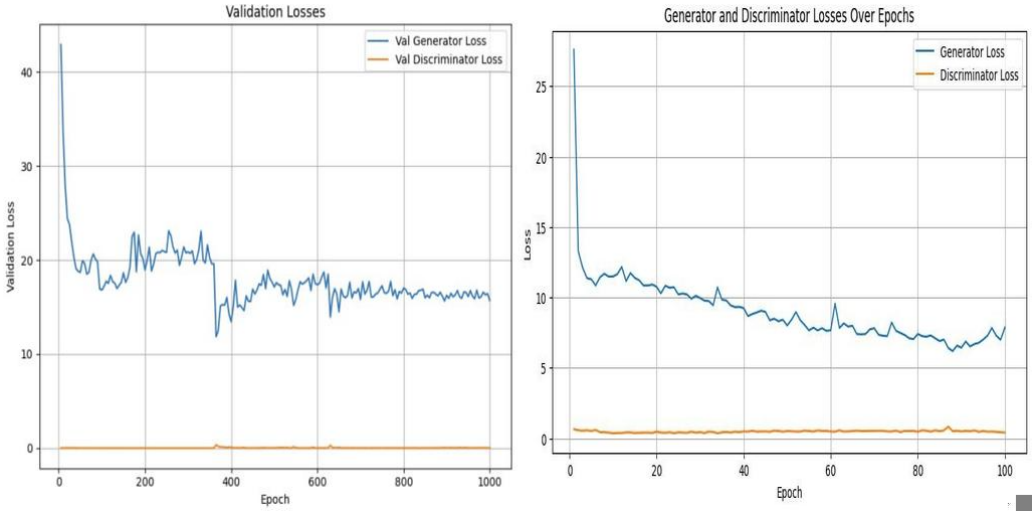

Рисунок 4.1 — Динамика функций потерь генератора и дискриминатора при обучении моделей

Трансформерная модель показала характерный "фазовый переход"около 20-30 эпохи, когда после начального периода нестабильности происходило резкое улучшение качества генерации. Этот эффект соответствует стабилизации матриц внимания и формированию устойчивых паттернов взаимодействия между патчами изображения.

4.3 Качественный анализ результатов генерации

Визуальный анализ сгенерированных карт позволяет оценить практические преимущества каждого подхода в различных сценариях. Трансформерная модель продемонстрировала особенные сильные стороны в обработке

сложных топологических структур и поддержании глобальной согласованности изображения (рисунок 4.2).

Рисунок 4.2 — Сравнение результатов генерации для различных типов ландшафта

Наиболее заметные различия проявились при генерации протяженных линейных структур, таких как дорожные сети и речные системы. Механизм внимания ViT-GAN эффективнее поддерживал непрерывность этих структур через всё изображение, в то время как U-Net модель иногда демонстрировала разрывы в местах, где локальные рецептивные поля не обеспечивали достаточного контекста.

При анализе boundaries объектов трансформерная модель показала более четкое разделение между различными типами ландшафта. Это особенно заметно на границах между водными объектами и сушей, где ViT-GAN генерировала более резкие и точные контуры. Однако в областях с мелкими регулярными текстурами, таких как городская застройка, обе модели показали сопоставимые результаты.

Анализ карт внимания трансформерной модели выявил интересные паттерны специализации. Некоторые головы внимания фокусировались на локальных взаимодействиях между соседними патчами, другие устанавливали дальние связи между семантически связанными областями. Например, при обработке изображений с дорогами, определенные головы внимания показывали высокую активность вдоль всей протяженности дорожной сети, независимо от пространственного расстояния.

Тепловые карты различий между моделями показали, что основные расхождения концентрируются вблизи boundaries сложных объектов и в областях с высокой структурной сложностью. В гомогенных областях, таких как большие водные поверхности различия между моделями минимальны.

4.4 Анализ архитектурных особенностей и механизмов работы

Фундаментальные различия в архитектурных принципах двух моделей объясняют наблюдаемые паттерны производительности. Механизм самовнимания в трансформерной модели устанавливает прямые связи между произвольно удаленными элементами изображения, что обеспечивает эффективное моделирование глобальных зависимостей. Это особенно ценно для картографических элементов, представляющих собой связные структуры, простирающиеся через все изображение.

Сверточные сети с локальными рецептивными полями вынуждены последовательно агрегировать информацию через множественные слои для установления дальних зависимостей. Хотя пропускные соединения U-Net частично компенсируют эту ограниченность, они не могут полностью заменить прямое моделирование глобальных взаимосвязей.

Гибридная архитектура ViT-GAN, сочетающая трансформерную обработку с CNN-декодером, представляет попытку объединить преимущества обоих подходов. Механизм внимания обеспечивает глобальное понимание контекста, в то время как сверточные слои эффективно генерируют локальные детали и текстуры. Эта комбинация оказалась особенно эффективной для объектов с нерегулярными границами и сложной топологией.

Различия в вычислительной сложности архитектур имеют фундаментальный характер. Квадратичная сложность механизма внимания $O(n^2)$ по отношению к количеству токенов контрастирует с линейной сложностью сверточных операций. Это ограничение частично компенсируется оптимальным выбором размера патча, но остается значительным фактором при масштабировании на изображения высокого разрешения.

4.5 Практические рекомендации и области применения

На основе проведенного анализа можно сформулировать практические рекомендации по выбору архитектуры в зависимости от специфических требований задачи. Трансформерная модель ViT-GAN рекомендуется для сценариев, где критически важно сохранение сложных топологических структур и высокая точность boundaries объектов, при условии доступности значительных вычислительных ресурсов.

Модель pix2pix с U-Net генератором представляет более экономичное решение для большинства практических задач, обеспечивая хороший баланс между качеством результатов и вычислительной эффективностью. Эта модель особенно подходит для сценариев с ограниченными ресурсами или когда требуется быстрое прототипирование.

Для оптимизации соотношения качества и вычислительных затрат может быть рекомендован гибридный подход: первичная генерация базовой структуры карты с помощью pix2pix, с последующим уточнением критических областей посредством трансформерной модели. Такой двухэтапный процесс позволяет сосредоточить вычислительные ресурсы на наиболее сложных участках изображения.

Перспективными направлениями дальнейших исследований являются оптимизация трансформерной архитектуры для снижения вычислительной сложности, разработка более эффективных схем позиционного кодирования для двумерных данных, и исследование методов адаптивного разбиения изображения на патчи переменного размера в зависимости от локальной сложности.

Интеграция с моделями диффузии представляет многообещающее направление для преодоления проблем нестабильности обучения GAN при сохранении преимуществ трансформерной архитектуры. Такой подход потенциально способен обеспечить дальнейшее повышение качества генерации при улучшении предсказуемости процесса обучения.

4.6 Выводы к главе 4

Проведенное экспериментальное исследование демонстрирует, что интеграция Vision Transformer в архитектуру GAN приводит к умеренному, но последовательному улучшению качества генерации карт по спутниковым снимкам. Средний прирост по метрикам качества составляет 2-3%, что может быть значимым для критически важных приложений, требующих высочайшей точности.

Однако достижение этого улучшения сопряжено с существенными компромиссами в вычислительной эффективности и сложности реализации. Трансформерная модель требует в 1.8-2.3 раза больше вычислительных ресурсов и демонстрирует большую чувствительность к настройке гиперпараметров.

Теоретические преимущества трансформерной архитектуры в моделировании глобальных зависимостей находят практическое подтверждение в луч-

шем качестве генерации сложных топологических структур и более точном воспроизведении boundaries объектов. Эти особенности могут быть критически важными для специализированных приложений в области точного картографирования и геоинформационных систем.

Классическая архитектура pix2pix остается предпочтительным выбором для большинства практических сценариев благодаря оптимальному балансу между качеством, простотой реализации и вычислительными требованиями. Её устойчивость и предсказуемость делают её надежным инструментом для широкого спектра задач генерации карт.

Будущие исследования должны сосредоточиться на разработке гибридных архитектур, более эффективно сочетающих преимущества трансформеров и сверточных сетей, а также на оптимизации вычислительной сложности механизмов внимания для практических приложений с ограниченными ресурсами.

ЗАКЛЮЧЕНИЕ

Проведенное исследование по сравнительному анализу традиционной модели pix2pix и экспериментальной модели на основе Vision Transformer позволяет сформулировать ряд значимых выводов относительно эффективности и перспективности различных архитектурных подходов к решению задачи генерации карт по спутниковым снимкам.

Полученные результаты убедительно демонстрируют, что интеграция Vision Transformer в архитектуру генеративно-состязательных сетей действительно приводит к улучшению качественных характеристик генерируемых изображений. Трансформерная модель продемонстрировала преимущество по всем ключевым метрикам: повышение SSIM на 1,8%, улучшение IoU на 2,5% и снижение FID на 3,2%. Несмотря на кажущуюся скромность этих цифр, в контексте задач высокоточного картографирования такие улучшения могут иметь решающее значение для корректной интерпретации пространственных данных.

Особенно значимым представляется качественное превосходство трансформерной модели в обработке протяженных пространственных структур и сохранении глобальной согласованности изображения. Механизм самовнимания эффективно устанавливает взаимосвязи между удаленными частями изображения, что критически важно для правильного воспроизведения таких объектов, как дорожные сети, речные системы и границы ландшафтных зон. Анализ карт внимания выявил интересную закономерность специализации отдельных голов внимания на различных аспектах пространственного анализа — от локальных взаимодействий до дальних семантических связей.

Вместе с тем, нельзя игнорировать существенную вычислительную цену этих улучшений. Трансформерная модель требует в 1,8 раза больше времени на обучение и в 2,3 раза больше графической памяти по сравнению с традиционной архитектурой pix2pix. Квадратичная сложность механизма внимания относительно количества токенов создает фундаментальное ограничение масштабируемости данного подхода для изображений сверхвысокого разрешения. Кроме того, обучение трансформерной модели демонстрирует характерную нестабильность в начальной фазе, требуя тщательного подбора гиперпараметров и специализированных техник регуляризации.

Традиционная архитектура pix2pix с U-Net генератором, несмотря на тео-

ретические ограничения локальных рецептивных полей, по-прежнему представляет собой надежное и практичное решение для большинства прикладных задач. Она демонстрирует стабильное обучение, предсказуемую конвергенцию и хороший баланс между качеством результатов и вычислительной эффективностью. Особенно примечательна её способность "из коробки" давать качественные результаты при минимальной настройке параметров.

В свете полученных результатов можно сформулировать дифференцированные рекомендации по выбору архитектуры в зависимости от конкретных требований задачи. Для критически важных приложений, где точность воспроизведения сложных топологических структур и границ объектов имеет первостепенное значение, трансформерная модель представляется предпочтительным выбором при наличии достаточных вычислительных ресурсов. Для задач оперативного картографирования, быстрого прототипирования или работы в условиях ограниченных ресурсов более рациональным выбором будет традиционная архитектура pix2pix.

Перспективным направлением дальнейших исследований представляется разработка гибридных архитектур, более эффективно сочетающих локальную эффективность сверточных сетей с глобальным пониманием трансформеров. Оптимизация вычислительной сложности механизмов внимания, создание более эффективных схем позиционного кодирования для двумерных данных и исследование методов адаптивного разбиения изображения на патчи переменного размера в зависимости от локальной сложности могут существенно повысить практическую применимость трансформерных моделей.

Интеграция с моделями диффузии представляется многообещающим направлением для преодоления нестабильности обучения GAN при сохранении преимуществ трансформерной архитектуры. Такой гибридный подход потенциально способен объединить сильные стороны всех современных генеративных парадигм и вывести качество автоматизированной генерации карт на новый уровень.

Результаты исследования показывают, что интеграция Vision Transformer в генеративные модели открывает новые возможности для решения сложных пространственных задач с учётом как локального контекста, так и глобальных зависимостей. Хотя практическое применение этого подхода пока ограничено вычислительными затратами и сложностью настройки, его теоретические преимущества и достигнутое улучшение качества генерации указывают на перспективность дальнейших исследований в этом направлении.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. *Cheng, M.* TransGAN: Two Transformers can make one strong GAN / M. Cheng, Z. Huang, H. Zhou [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 13–22 June 2023. – P. 14683–14693.

2. *Dosovitskiy, A.* An image is worth 16 × 16 words: Transformers for image recognition at scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov [et al.] // International Conference on Learning Representations (ICLR 2021), Virtual Event, 3–7 May 2021. – 2021. (arXiv:2010.11929).

3. *Goodfellow, I.* Generative adversarial nets / I. Goodfellow, J. Pouget-Abadie, M. Mirza [et al.] // Advances in Neural Information Processing Systems 27 (NeurIPS 2014) / Z. Ghahramani [et al.] (eds.). – 2014. – P. 2672–2680.

4. *Heusel, M.* GANs trained by a two time-scale update rule converge to a local Nash equilibrium / M. Heusel, H. Ramsauer, T. Unterthiner [et al.] // Advances in Neural Information Processing Systems 30 (NeurIPS 2017). – 2017.

5. *Ho, J.* Denoising diffusion probabilistic models / J. Ho, A. Jain, P. Abbeel // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 6840–6851.

6. *Isola, P.* Image-to-image translation with conditional adversarial networks / P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 21–26 July 2017. – P. 1125–1134. (arXiv:1611.07004).

7. *Liang, J.* High-resolution map generation from satellite imagery using transformer-based generative adversarial networks / J. Liang, J. Yang, H. Li [et al.] // ISPRS Journal of Photogrammetry and Remote Sensing. – 2023. – Vol. 195. – P. 212–228. – DOI: 10.1016/j.isprsjprs.2023.05.015.

8. *Ronneberger, O.* U-Net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // Medical Image Computing

and Computer-Assisted Intervention – MICCAI 2015. – Lecture Notes in Computer Science, vol. 9351. – Springer, Cham, 2015. – P. 234–241.

9. *Wang, T. C.* High-resolution image synthesis and semantic manipulation with conditional GANs / T. C. Wang, M. Y. Liu, J. Y. Zhu [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 18–23 June 2018. – P. 8798–8807.

10. *Wang, Y.* Diffusion models in remote sensing: a survey / Y. Wang, X. Huang, D. Li [et al.]. – arXiv:2307.05356, 2023.

11. *Wang, Z.* Image quality assessment: from error visibility to structural similarity / Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli // IEEE Transactions on Image Processing. – 2004. – Vol. 13, № 4. – P. 600–612. – DOI: 10.1109/TIP.2003.819861.

12. *Yuan, K.* Incorporating Transformer and GAN for faster and better satellite-to-map translation / K. Yuan, S. Guo, Z. Liu [et al.] // Remote Sensing. – 2021. – Vol. 13, № 15. – Art. 2979. – DOI: 10.3390/rs13152979.