ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

На правах рукописи

Бережнов Никита Игоревич

СОВЕРШЕНСТВОВАНИЕ МЕХАНИЗМОВ ВНИМАНИЯ В ГЛУБОКИХ НЕЙРОННЫХ СЕТЯХ – ТРАНСФОРМЕРАХ В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ И АУГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Специальность 1.2.1. Искусственный интеллект и машинное обучение

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель: А. А. Сирота
доктор технических наук, профессор

Воронеж – 2025

Содержание

Введение

Актуальность темы диссертации. Современные системы компьютерного зрения предъявляют высокие требования к качеству визуальной информации, получаемой при обработке изображений в автоматическом режиме. В реальных условиях изображения, поступающие на вход таких систем, часто оказываются искаженными из-за наличия шумов, влияния погодных осадков, технических ограничений сенсоров, случайных помех или ошибок передачи данных. Решение задач восстановления и улучшения качества изображений, а также эффективной аугментации обучающих данных становится особенно важным в контексте роста сложности прикладных задач и разнообразия реальных условий съемки.

В последние годы активно развиваются методы глубокого обучения, демонстрирующие высокую устойчивость получаемых при их использовании результатов к сложным типам искажений. Среди них большую роль играют архитектуры глубоких нейронных сетей (ГНС) трансформерного типа, успешно адаптированные для задач компьютерного зрения. Механизм внимания (attention), лежащий в их основе, позволяет эффективно учитывать как локальные, так и глобальные взаимосвязи между элементами изображения, что важно при устранении артефактов, зашумления, затенения или других отклонений от исходного вида изображения.

Однако практическое применение трансформеров в задачах восстановления изображений часто сопряжено с рядом ограничений. Во-первых, внимание может избыточно фокусироваться на отдельных структурах изображений, что снижает эффективность восстановления от шумов и искажений. Во-вторых, большая глубина архитектур делает их склонными к переобучению, особенно при ограниченных объемах обучающих данных. В-третьих, известные трансформерные модели, используемые в задачах обработки изображений, весьма громоздки и требуют больших вычислительных ресурсов (памяти, времени), особенно на этапе обучения. В-четвертых, сбор достаточного объема разнообразных изображений, необходимых для обучения в реальных условиях

затруднен, что делает задачу формирования качественного обучающего датасета нетривиальной.

Особенно критична ситуация с получением изображений объектов в условиях различных погодных осадков, затрудняющих восприятие сцен. С этой целью возможно использовать методы аугментации – искусственного размножения данных, в которых реализуются различные способы изменения наборов обучающих данных применительно к различным внешним условиям.

Решение проблемы нехватки данных требует применения как классических приемов аугментации, так и генеративных нейросетевых моделей, способных формировать новые условно-реалистичные изображения. При этом важно учитывать не только разнообразие, но и статистическую достоверность синтезируемых данных. Особенно актуален этот подход в задачах, связанных с нечеткими, нелинейными или аппликативными видами шумов (например, помехи на медицинских изображениях или сигналы с электронных микроскопов), где традиционные методы фильтрации оказываются неэффективными.

Применение генеративных нейросетей (GAN, VAE, диффузионные модели, модели-трансформеры) позволяет формировать синтетические изображения, имитирующие сложные искажения: атмосферные осадки, цифровые дефекты, шумы различных типов. Это открывает возможности как для создания обширных обучающих наборов данных, так и для реализации архитектур, в которых восстановление и аугментация изображений выступают как взаимосвязанные стадии единого процесса обработки информации. В частности, генерация зашумленных или стилизованных изображений может быть непосредственно добавлена в этапы обучения нейросетевых моделей, обеспечивая лучшее понимание вариативности входных данных.

Однако подобные подходы требуют более тщательной регуляризации и контроля процесса обучения. В противном случае возникают риски искажения исходного распределения данных, доменного смещения данных (bias) и ухудшения обобщающей способности модели. Кроме того, существующие генеративные модели имеют сложности в процессе обучения и склонность к

неконтролируемому синтезу изображений. Это требует разработки гибридных архитектур и дальнейшего совершенствования механизмов внимания с учетом специфики решаемой задачи.

Таким образом, актуальными являются исследования комбинированных подходов для решения задачи восстановления изображений, сочетающих относительно «легковесные» архитектуры трансформеров с модифицированными механизмами канального и пространственного внимания и современные методы аугментации данных, обеспечивающие увеличение устойчивости, точности и обобщающей способности обучаемых нейросетевых моделей. Такие решения позволяют не только обеспечить качественное восстановление изображений, но и повысить устойчивость систем компьютерного зрения в реальных и, зачастую, нестабильных условиях.

Степень разработанности темы диссертации. Алгоритмы восстановления и аугментации изображений достаточно давно развиваются в области компьютерного зрения и машинного обучения. Классические подходы, основанные на фильтрации, частотных преобразованиях и методах регуляризации, подробно рассматривались в трудах Н.Н. Бондиной, В.С. Сизикова, В.П. Кузнецова, В.К. Клочко, Р.С. Гонсалеса, С.О. Емельянова, В.И. Тихонова и др. [1-11]. Однако подобные методы часто оказываются неэффективными при обработке изображений, содержащих сложные и нелинейные искажения, такие как шумы нестандартной природы, атмосферные эффекты или структурные дефекты.

Существенный прогресс в решении этих задач достигнут за счет использования методов глубокого обучения, в частности, сверточных нейронных сетей (CNN) [21-26] и архитектур автокодировщиков [13, 16]. Однако основные успехи здесь связаны с появлением трансформеров, реализующих принципы механизмов внимания (самовнимания), адаптированных к задачам обработки изображений. Архитектуры Vision Transformer (ViT), Swin Transformer, Restormer и их производные, предложенные в работах A. Dosovitskiy, Z. Liu, H. Zhao, S.W. Zamir, J. Valanarasu, P. Isola и других [12, 27-36], продемонстрировали высокую

эффективность в задачах восстановления изображений, улучшения визуального качества и удаления шумов.

Развитие механизмов внимания и их модификаций на основе структурной регуляризации и стохастического сглаживания подробно рассматривались в исследованиях A. Vaswani, B. Li, W. Zhou, H. Lu, X. Chen, K. He и др. [9, 90–93]. Тем не менее, несмотря на достигнутый прогресс, актуальной остается проблема переобучения и повышения эффективности моделей трансформерного типа, особенно в условиях недостаточности обучающих выборок.

Для решения данной проблемы используются также методы аугментации изображений. Наряду с эвристическими подходами, такими как RandAugment и AutoAugment [48, 49], все более широкое распространение получают генеративные методы, включая модели класса GAN, VAE, диффузионные модели и трансформеры с перекрестным вниманием [37-39, 50, 55, 59-61]. Они позволяют создавать синтетические изображения с реалистичными атмосферными искажениями, шумами и стилем, что существенно расширяет возможности моделирования входных данных. Особый интерес представляют модели WeatherDG, RainDiffusion и TransWeather [63-69], ориентированные на синтез осадков и погодных эффектов.

Отдельного внимания заслуживают работы, направленные на визуализацию и интерпретацию внимания в трансформерах [83-85], а также исследования по применению регуляризации механизма внимания для повышения устойчивости нейросетевых моделей к шумам и повышению их обобщающей способности [91-93].

Тем не менее, несмотря на большое количество исследований, остаются нерешенными многие задачи интеграции модифицированных механизмов внимания и алгоритмов генерации синтетических искажений в единую систему обработки, способную одновременно восстанавливать изображения и эффективно увеличивать набор обучающих данных. Актуальной задачей также остается разработка малозатратных архитектур для обработки изображений, обеспечивающих качество, сопоставимое с лучшими известными большими

моделями, и устойчивых к проблемам, возникающим при несбалансированной генерации и некорректной стилизации изображений.

Таким образом, настоящая работа продолжает и развивает научные исследования, направленные на повышение эффективности нейросетевых архитектур восстановления изображений на основе сетей трансформерного типа с модифицированными механизмами внимания и аугментацией обучающих данных для создания реалистических изображений в условиях погодных искажений и воздействия различных шумов. Работа находится в русле актуальных направлений исследований в области искусственного интеллекта и машинного обучения.

Цель и задачи исследования. Целью диссертационной работы является совершенствование алгоритмов восстановления изображений в условиях различных искажений (включая шумы, атмосферные осадки, аппликативные помехи) на основе архитектур ГНС трансформерного типа с модифицированными механизмами внимания и применение средств аугментации данных, направленных на повышение обобщающей способности нейронных сетей, достигаемой в процессе обучения.

Для достижения указанной цели в работе решаются следующие задачи:

1. Анализ современных подходов к восстановлению и аугментации изображений, выявление ограничений существующих алгоритмов и обоснование необходимости внедрения новых архитектурных решений в моделях глубокого обучения.

2. Теоретическое обоснование и разработка методов модификации и регуляризации механизма внимания в трансформерах, направленных на повышение их устойчивости к переобучению и улучшение качества восстанавливаемых изображений.

3. Синтез архитектур трансформеров с усовершенствованными канальным и пространственным механизмами внимания для восстановления изображений в условиях сложных помех и искажений.

4. Разработка алгоритмов генерации и стилизации изображений на основе ГНС в целях аугментации обучающих данных, включая синтез реалистичных изображений объектов в условиях атмосферных осадков.

5. Разработка программного комплекса и методики его применения, реализующего комплексное применение предложенных алгоритмов восстановления и аугментации изображений, проведение экспериментального анализа для выявления их эффективности в типовых задачах компьютерного зрения.

Объект исследования. Объектом исследования являются системы компьютерного зрения для восстановления и аугментации изображений.

Предмет исследования. Предметом исследования являются модели и алгоритмы глубокого обучения на основе трансформеров с модифицированными механизмами внимания, а также алгоритмы генерации и стилизации условно-реальных изображений, используемые для повышения качества изображений и увеличения обучающих выборок в задачах компьютерного зрения.

Методы исследования. В ходе выполнения диссертационной работы использовались методы математического анализа, линейной алгебры и оптимизации, методы теории вероятностей, методы цифровой обработки изображений, модели и методы глубокого машинного обучения, технологии разработки многослойных ГНС (в том числе, сверточных и трансформерных архитектур), модели и методы генерации и стилизации изображений, методы и средства имитационного моделирования, технологии программирования ГНС с использованием современных инструментальных сред.

Научная новизна диссертации заключается в следующем.

1. Предложен и теоретически обоснован способ структурной регуляризации механизма внимания в нейронных сетях трансформерного типа, отличающийся использованием мультипликативной и аддитивной стохастической составляющих, вносимых при вычислении матриц весов внимания, что обеспечивает сглаживание распределения весов для предотвращения их неконтролируемого роста в процессе обучения.

2. Предложен и теоретически обоснован способ структурной регуляризации процесса обучения трансформеров, отличающийся использованием обучаемой матрицы масштабных коэффициентов, что позволяет оказывать позитивное влияние в ситуациях насыщения активационной функции механизма внимания.

3. Разработаны и исследованы модификации канального механизмов внимания в трансформерах, отличающиеся использованием сжатия канальных признаков. Предложены способы структурной регуляризации пространственного внимания. Это позволило повысить качество восстановления изображений при одновременном снижении вычислительной сложности моделей. На основе этого предложены улучшенные архитектуры нейронных сетей трансформерного типа. Проведены экспериментальные исследования, подтверждающие возникновение положительного эффекта в задачах восстановления изображений и улучшение значений метрик качества по сравнению с базовыми прототипами.

4. Разработаны модели и алгоритмы аугментации изображений на основе специализированных архитектур нейронных сетей. Особое внимание уделено синтезированию изображений объектов в условиях атмосферных осадков, затрудняющих восприятие анализируемых сцен. Предложена новая архитектура модели трансформер, объединяющая сверточный энкодер-декодер и перекрестный механизм внимания для генерации атмосферных осадков (дождь, снег, туман), позволяющая сохранить структурную целостность сцены. Введена составная функция потерь для обучения предложенной модели в условиях различных погодных осадков, учитывающая различные аспекты качества синтезирования изображений.

5. Разработан программный комплекс алгоритмов восстановления и аугментации изображений, основанный на объединении архитектур трансформеров с усовершенствованными механизмами внимания и специализированных моделей аугментации изображений и синтеза искажений. Предложена методика их совместного применения для повышения устойчивости нейросетевых моделей к различным помехам в условиях нехватки обучающих

данных. Экспериментально подтверждена эффективность использования синтезированных изображений в качестве обучающих данных в задачах восстановления, классификации и сегментации.

Тематика работы полностью соответствует паспорту специальности 1.2.1. Искусственный интеллект и машинное обучение по пунктам:

п.4. Разработка методов, алгоритмов и создание систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке, для изображений, речи, биомедицины и других специальных видов данных;

п.14. Методы и средства формирования массивов условно-реальных данных и прецедентов, необходимых для решения задач искусственного интеллекта и машинного обучения.

Теоретическая и практическая значимость. Теоретическая значимость работы заключается в развитии подходов к совершенствованию архитектур ГНС трансформерного типа, направленных на решение задач восстановления изображений, на основе структурной регуляризации механизмов внимания. Проведенные в этом плане теоретические обоснования и доказательства носят достаточно общий характер и могут быть использованы для построения ГНС с различными вариантами реализации механизма внимания при решении других задач, в том числе, задач классификации и семантической сегментации. Предложенные модели и алгоритмы аугментации позволяют повысить обобщающую способность моделей в условиях ограниченности данных, наличия сложных искажений и помех для различных задач компьютерного зрения.

Представленные теоретические и экспериментальные результаты позволяют проводить сравнительный анализ альтернативных подходов к построению алгоритмов обработки информации рассматриваемого класса и выбор конкретного алгоритма с учетом возникающих на практике ограничений.

Практическая значимость обусловлена возможностью внедрения разработанных алгоритмов в прикладные системы компьютерного зрения, включая: автоматические системы видеонаблюдения и мониторинга;

аэрокосмическую съемку в сложных погодных условиях; медицинскую томографию (для подавления шумов и артефактов на снимках); системы обработки изображений в мобильных устройствах.

Алгоритмы аугментации, в частности, предложенная в работе архитектура WeatherTransformer позволяют генерировать синтетические обучающие данные, моделирующие реальные условия съемки без необходимости их ручного сбора, что особенно актуально при обучении нейронных сетей на малых или несбалансированных выборках данных.

Результаты работы использованы при выполнении в ФГБОУ ВО «Воронежский государственный университет» научно-исследовательских работ в период 2022-2025 годов, связанных с обработкой изображений специального назначения (НИЧ-21009, НИЧ-23019), в которых автор являлся непосредственным исполнителем, а также в учебном процессе вуза.

Положения и результаты, выносимые на защиту. На защиту выносятся следующие результаты и положения.

1. Структурная регуляризация механизма внимания в трансформерных блоках нейронных сетей может осуществляться путем внесения мультипликативной или эквивалентной ей аддитивной стохастической составляющей при вычислении весов внимания, что проявляется в сглаживании соотношения весов внимания для снижения возможности их неконтролируемого роста в процессе обучения.

2. Применение отдельно обучаемой матрицы масштабных коэффициентов в качестве мультипликативной составляющей при вычислении матриц внимания в рамках стандартного механизма позволяет снижать воздействие возникающих аномалий в виде существенно превалирующих весов внимания, ситуаций насыщения активационной функции и включает дополнительные возможности регулирования весовых коэффициентов внимания.

3. Повышение качества восстановления изображений в стандартных архитектурах глубоких нейронных сетей может быть достигнуто на основе предложенных модификаций, реализующих добавление аддитивной

стохастической составляющей в виде выборочных оценок дисперсионных характеристик признаков, вычисляемых в предшествующих сверточных слоях, и обучаемых матриц масштабных коэффициентов.

4. Снижение вычислительной сложности модулей внимания в трансформерах без существенных потерь качества восстановления изображений может быть достигнуто за счет использования предложенного алгоритма канального сжатия, что позволяет эффективно учитывать как пространственные, так и канальные зависимости, а также масштабировать архитектуру сети для входных изображений высокого разрешения.

5. Эффективный и малозатратный подход к аугментации изображений с целью учета факторов, негативных для восприятия сцен и, прежде всего, атмосферных осадков (дождь, снег, туман), может быть достигнут за счет их переноса из эталонного и включения в модифицируемое изображение на основе использования предложенной модели двухвходового трансформера, объединяющего сверточный энкодер-декодер и перекрестный механизм внимания.

Степень достоверности результатов работы. Результаты исследований, сформулированные в диссертационной работе, основаны на теоретических и экспериментальных методах исследований, взаимно дополняющих друг друга, и согласуются между собой. Указанные результаты получены с использованием комплекса теоретических, вычислительных и экспериментальных методов. Проведенные исследования основываются на строго формализованных постановках задач, апробированных подходах глубокого обучения и алгоритмах обработки изображений. Все разработанные модели, механизмы внимания и алгоритмы аугментации подвергались тестированию в контролируемых условиях на синтетических и реальных датасетах с использованием общепринятых метрик качества.

Корректность и воспроизводимость синтезированных архитектур трансформеров, модификаций внимания и генеративных моделей подтверждается результатами многочисленных вычислительных экспериментов, сопоставлением с

базовыми методами, а также статистической обработкой результатов. Выводы, сделанные в работе, имеют обоснованную интерпретацию, совпадают в ряде частных случаев с результатами, полученными другими авторами, и согласуются с общепринятыми теориями в области машинного обучения и обработки изображений. Таким образом, полученные в ходе диссертационной работы результаты можно считать в достаточной степени обоснованными, достоверными и практически значимыми.

Апробация работы. Основные положения, выводы и рекомендации, сформулированные в диссертации, докладывались и обсуждались на ряде научных конференций различного уровня. В частности, результаты были представлены:

- на XXIII, XXIV и XXV Международных конференциях «Информатика: проблемы, методология, технологии» (г. Воронеж, 2023–2025 гг.);
- на 5-й Международной конференции «International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)» (г. Липецк, 2023 г.);
- на ежегодных научных сессиях факультета компьютерных наук ВГУ.

Публикации. По теме диссертационной работы опубликовано 8 научных работ, из них 4 статьи в изданиях, рекомендованных ВАК и 1 статья в материалах конференции, представленной в IEEE Explore (Scopus), получено 1 свидетельство о государственной регистрации программы для ЭВМ.

Все выносимые на защиту результаты и положения принадлежат лично автору. В публикациях, выполненных в соавторстве с руководителем, последнему принадлежат постановка задачи и выбор направления исследований. Непосредственно соискателю принадлежат: обоснование предложенных архитектур и моделей, разработка алгоритмов, реализация программных прототипов, постановка и проведение экспериментов, анализ и интерпретация результатов.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы. Объем работы составляет 161

страницу основного текста, включая два приложения, 28 рисунков и 16 таблиц. Список использованных источников содержит 108 наименований.

В первой главе проводится всесторонний анализ современных методов и алгоритмов восстановления и аугментации изображений. Рассматриваются как классические подходы к восстановлению (линейная и нелинейная фильтрация, частотные преобразования), так и современные нейросетевые методы, включая сверточные архитектуры и трансформеры. Значительное внимание уделено алгоритмам аугментации: как эвристическим, так и генеративным моделям (GAN, VAE, диффузионным моделям). Особо выделяется направление синтеза условно-реальных изображений с погодными эффектами. В завершение главы формулируется общая схема построения проведения исследований в интересах создания новых моделей и алгоритмов восстановления и аугментации изображений.

Во второй главе обосновываются и исследуются возможные способы модификации механизма внимания в трансформерах для задач восстановления изображений. Подробно рассматриваются особенности вычисления механизма внимания (самовнимания) и его взаимосвязь с различными искажениями изображения. Предлагаются и теоретически обосновываются способы регуляризации механизма внимания в процессе обучения.

В третьей главе осуществляется синтез и исследование алгоритмов восстановления изображений на основе моделей-трансформеров с модифицированными механизмами внимания. Предлагаются и исследуются две архитектуры: одна со сжатием канального механизмом внимания, другая с модифицированным пространственным. Приводятся результаты экспериментов, подтверждающих преимущество предложенных моделей по сравнению с базовыми архитектурами по метрикам пикового отношения сигнал-шум и меры структурного сходства изображений, а также с точки зрения вычислительной сложности процесса обучения.

В четвертой главе рассматриваются алгоритмы генерации и стилизации изображений для целей аугментации. Приводится систематизация возможных

подходов: от простых эвристик до сложных генеративных архитектур. Особое внимание уделено предложенной новой модели *WeatherTransformer*, реализующей синтез атмосферных осадков с использованием двухвходовой модели-трансформера и перекрестного внимания. В заключение главы описана методика применения разработанных алгоритмов в задачах сегментации, классификации и восстановления изображений, приводится структура программного комплекса, реализующего предложенные модели и алгоритмы.

1. Анализ существующих методов и алгоритмов восстановления и аугментации изображений. Общая схема исследования

В настоящей работе используются понятия восстановление изображений (image restoration), а также улучшение качества изображений (image enhancement). Восстановление направлено на устранение любого рода искажений, возникших в процессе получения (регистрации) изображения с целью приблизить изображение к его исходному, «чистому» виду. При этом обычно используется количественный критерий близости (метрика). В задаче улучшения качества изображений исследователи фокусируются в большей степени на субъективном улучшении визуального восприятия изображения, повышении его информативности. Хотя эти задачи и близко связаны, но, тем не менее, не всегда эквивалентны. Далее в качестве основной будет рассматриваться задача восстановления изображений (ВИ).

Шумы и искажения на изображениях могут возникать как на этапе обработки, так и при передаче и хранении данных. К основным источникам шумов и искажений относятся: аппаратные ограничения (шумы сенсоров видеокамер, ТВ-тюнеров, сканеров), неблагоприятные условия съемки, например, низкая освещенность, атмосферные осадки, электромагнитные помехи в каналах связи и передачи данных изображений, повреждения датчиков и носителей, а также артефакты при декодировании видеосигналов. Особый случай представляет собой спекл-шум, возникающий в когерентных системах (радиолокация, УЗИ) в результате интерференции отраженных волн от мелких неоднородностей.

По характеру воздействия шум может быть аддитивным, мультипликативным или импульсным (аппликативным, замещающим). Различают белый шум (некоррелированные значения), гауссовский шум с пространственной корреляцией, импульсный шум («соль и перец»), цветовые артефакты, а также цифровой шум, проявляющийся в виде случайных флуктуаций яркости и цвета отдельных пикселей. Последний особенно заметен на равномерных или размытых

областях изображений, снижая визуальное качество и ухудшая восприятие деталей.

Шумы и искажения различаются по природе, интенсивности, масштабу воздействия на изображение. Для их описания используются различные математические модели. Например, в работе [1] авторы изучают шумы на изображениях, полученных при помощи электронного микроскопа. Они приходят к выводу, что данный шум нельзя аппроксимировать ни гауссовским, ни пуассоновским распределением. Он имеет более сложную природу, поэтому необходимо использовать другие, более сложные модели для восстановления изображений при наличии такого шума.

В общем случае решение задачи восстановления искомого изображения в присутствии аддитивного шума может быть найдено путем решения интегрального уравнения Фредгольма первого рода:

$$z(x,y)=\int_{a_1}^{b_1}\int_{a_2}^{b_2} K(x,s_1,y,s_2)f(s_1,s_2)ds_1ds_2, \quad a_1\le x\le b_1,\ a_2\le y\le b_2,$$

где $f(s_1,s_2)$ – исходное неискаженное изображение; $K(x,s_1,y,s_2)$ – функция ядра, описывающая преобразование исходного изображения; $z(x,y)$ – результат искажения**.**

Найти точное решение уравнения на практике часто не представляется возможным, так как задача является обратной некорректно поставленной по Адамару из-за её неустойчивости [2]. Малые отклонения в данных $z(x,y)$, например, вызванные шумом, могут приводить к существенно различающимся решениям $f(s_1,s_2)$. Это связано с тем, что оператор, заданный ядром K, часто является плохо обусловленным или вырожденным.

Для получения устойчивого приближенного решения применяются методы регуляризации. К наиболее распространенным методам относятся:

– регуляризация по Тихонову – поиск функции $f(s_1,s_2)$, минимизирующей функционал:

$$J_\alpha[f] = \left\| z(x,y) - \iint K(x,s_1,y,s_2) f(s_1,s_2) ds_1 ds_2 \right\|^2 + \alpha \| Lf \|^2,$$

где $\alpha > 0$ – параметр регуляризации, а L – стабилизатор, который на практике делают оператором априорной гладкости (например, тождественный или лапласиан);

– методы оптимальной фильтрации такие, как, например, фильтр Калмана, который обеспечивают минимизацию среднеквадратичной ошибки на основе априорных статистических характеристик изображения и шума;

– итерационные и статистические методы, которые могут учитывать, как модель оператора K, так и вероятностные свойства шума.

Выбор конкретного метода не является универсальным и зависит от структуры ядра K, уровня шума, и доступной априорной информации об изображении $f(s_1,s_2)$. Поэтому на сегодняшний день исследователи все чаще обращаются к использованию глубоких нейронных сетей (ГНС), показывающих большую универсальность и обобщающую способность в задачах восстановления изображений в условиях сложных и плохо формализуемых моделей искажений.

При этом нейросетевые методы требуют больших объемов обучающих данных, а если этих данных недостаточно, то возможно переобучение, когда нейронная сеть приспосабливается к обучающими данным и на реальных данных дает худший результат, неприемлемый на практике. Отдельно исследователи отмечают предвзятость нейронных сетей (bias-проблема), выражающуюся в смещении получаемых на выходе результатов. Например, в работе [3] авторы показывают, что сверточные нейронные сети обладают структурной предвзятостью (structural bias), позволяющей эффективно восстанавливать изображения без использования больших наборов обучающих данных. Однако это же смещение часто приводит к переобучению и потере обобщающей способности нейронной сети, что показано автором в работе [4].

В задачах ВИ описанные выше проблемы также сохраняются: необходимо не только очистить изображение от шумов и искажений, но и сохранить его реалистичность и детали сцены. Однако из-за разнообразия типов и природы

шумов и искажений, исследователям часто трудно собрать реальный набор данных для обучения нейронной сети. Поэтому в дополнение используются техники аугментации – искусственного размножения данных, позволяющие увеличить набор обучающих данных для глубоких нейронных сетей путем синтеза и генерации новых изображений.

На сегодняшний день существует множество методов аугментации данных, использующих как эвристические алгоритмы, так и нейросетевые подходы. Например, авторы [5] генерируют зашумленные изображения при помощи генеративных состязательных сетей (GAN). Однако часто разнообразие генерируемых и синтезируемых данных ограничивается определенным конечным набором шаблонов и эвристических алгоритмов. В дальнейшем в работе исследуются современные алгоритмы восстановления изображения и показывается, как продвинутые техники аугментации данных способны существенно улучшить качество работы нейронных сетей.

Как уже отмечалось, в качестве основной в работе будет рассматриваться задача восстановления изображений (ВИ). Однако при решении задачи аугментации обучающих данных будут использоваться также аспекты улучшения качества изображений. Это связано с тем, что для дальнейшей постобработки аугментированных изображений в системах компьютерного зрения не обязательно требуется точность восстановления, при этом эффективность алгоритмов аугментации проверяется на основе метрик, полученных при решении целевых задач классификации, сегментации и обнаружения объектов на изображениях.

1.1. Анализ известных методов восстановления изображений

1.1.1. Классические алгоритмы восстановления изображений

В настоящее время в системах компьютерного зрения используются разнообразные методы обработки и улучшения изображений. Классические (эвристические) подходы практически все сводятся к реализации тех или иных алгоритмов линейной или нелинейной фильтрации (фильтры среднего

арифметического, фильтры среднего геометрического, фильтры среднего гармонического, фильтры, основанные на порядковых статистиках, медианные фильтры, линейные сглаживающие и подчеркивающие фильтры, нелинейные фильтры и т.д.). При этом возникает проблема их выбора в плане применимости к конкретным изображениям.

К сожалению, на сегодняшний день не существует универсального алгоритма обработки. Так, например, в работе [6] авторы приводят эффективный алгоритм восстановления текстурных изображений малого масштаба на основе метода градиентных гистограмм, но при этом показывают, что его применимость существенно ограничена. Кроме того, современные исследователи фокусируется на отдельных типах изображений и их помехах. Например, медианный фильтр хорошо борется с шумом «соль и перец», но делает изображение размытым. Наоборот, алгоритмы повышения резкости страдают от присутствия высокочастотных помех на изображениях, которые также усиливаются с увеличением резкости.

К числу наиболее распространенных относится математическая модель линейной фильтрации изображения, которая описывает дискретную операцию по отношению к локальной области пикселей и описывается выражением:

$$I_{\text{new}}(x,y)=\sum_{k=1}^{m}\sum_{l=1}^{n}A_{kl}\,I_{\text{old}}\left(x-\frac{m}{2}+k,y-\frac{n}{2}+l\right),$$

где $I_{\text{old}}(x,y)$ и $I_{\text{new}}(x,y)$ – значения пикселей изображения до и после фильтрации соответственно; A_{kl} – коэффициенты фильтра, определяющие весовое влияние соседних пикселей; параметры m и n задают размерность окна. Данный тип фильтрации относится к классу линейных локальных операций, обладающих свойством суперпозиции.

Если исходное изображение искажается за счет линейного инвариантного к сдвигу оператора, то задача восстановления может быть сведена к модели свертки. С математической точки зрения свертка эквивалентна модели линейной фильтрации, описанной выше, но для модели восстановления изображений

необходимо еще учесть аддитивный шум. В этом случае искаженное изображение $z(x,y)$ представляется как свертка истинного изображения $f(x,y)$ с функцией ядра (или функцией рассеяния точки) $h(x,y)$ и добавлением аддитивного шума $n(x,y)$:

$$z(x,\,y) \;=\; (h \;*\; f)(x,\,y) \;+\; n(x,\,y)$$

Поскольку операция свертки в пространственной области соответствует умножению в частотной области, то целесообразно перейти к представлению через преобразование Фурье. После применения двумерного преобразования Фурье к обеим частям уравнения свёртки, можем получить следующее уравнение:

$$Z(u,\,v) \;=\; H(u,\,v)\,F(u,\,v) \;+\; N(u,\,v),$$

где $Z(u,\,v)$, $F(u,\,v)$, $H(u,\,v)$, $N(u,\,v)$ – Фурье-образы функций искаженного изображения, исходного изображения, функции ядра и шума соответственно. Это представление удобно для реализации алгоритмов восстановления, так как позволяет применять частотные фильтры.

Одним из наиболее эффективных подходов к восстановлению в условиях шума является Винеровская фильтрация, основанная на минимизации среднеквадратичной ошибки между истинным и восстановленным изображением. Оптимальный Винеровский фильтр в частотной области определяется выражением:

$$\hat{F}(u,v) = \frac{1}{H(u,v)} \frac{\left|H(u,v)\right|^2 G(u,v)}{\left|H(u,v)\right|^2 + S_\eta(u,v) / S_f(u,v)}.$$

Функцией S здесь обозначены энергетические спектры шума и исходного изображения соответственно. На практике дробь $S_\eta(u,v) / S_f(u,v)$ заменяется на некоторую константу, которую можно охарактеризовать как отношение сигнал-шум.

Таким образом, линейная фильтрация, представленная в пространственной форме, может быть выражена в виде свертки, которая в частотной области позволяет применять оптимальные методы восстановления, в частности

Винеровскую фильтрацию, обладающую высокой устойчивостью к шуму и эффективно подавляющую искажения.

При восстановлении изображений фильтром Винера из-за неполной информации о сигнале могут возникать краевые эффекты, также он малопригоден для восстановления изображения в присутствии импульсных помех. Поэтому для устранения такого рода искажений часто используются разного рода эвристические и морфологические методы восстановления изображений.

Большинство из этих фильтров работает с определенными типами изображений. Им также необходима тонкая настройка параметров, которые можно эффективно подобрать только с помощью субъективной оценки качества изображения. Из-за этого в последнее время набирают популярность методы адаптивной фильтрации изображений [7], такие как: адаптация структуры фильтра; адаптация размера окна обработки; адаптация коэффициентов фильтра обработки; адаптация принципов обработки; адаптация типа фильтрации.

Как уже сказано выше, существует большое количество стандартных классических алгоритмов обработки изображений, применяемых с целью повышения их качества, использующих пространственную и частотную фильтрацию, адаптивные методы коррекции изображений, а также разнообразные нелинейные фильтры [8, 9]. Но они не являются универсальными и, в большинстве случаев, необходимы предварительные исследования для определения перечня применяемых алгоритмов и их параметров, что весьма субъективно из-за отсутствия истинных, не зашумленных изображений. В этом плане следует отметить работу [10], где авторы проводят восстановление изображений в микроволновом диапазоне длин волн и указывают на сложность проверки качества результата из-за отсутствия эталонов и тонкой настройки их параметров.

Таким образом, применимость классических алгоритмов ограничивается конкретной задачей с известной математической моделью шумов и искажений. В связи с этим, как уже упоминалось, популярность набирают нейросетевые

алгоритмы с использованием технологий глубокого обучения, которые реализуют универсальный подход к решению задачи практически любой сложности.

1.1.2. Алгоритмы восстановления изображений на основе глубоких нейронных сетей

При решении задачи ВИ с помощью нейронных сетей необходимо также сформулировать задачу оптимизации. Для этого представим интегральное уравнение Фредгольма первого рода в виде интегрального оператора вида [11]:

$$z = Af + n,$$

где f – тензор исходного неискаженного изображения; A – линейный или нелинейный интегральный оператор; n – аддитивный шум; z – результат искажения.

Задачу оптимизации тогда можно записать в следующем виде:

$$\hat{z} = \inf_f \rho(Af, z),$$

где ρ – метрическая функция; $\hat{z}$ – восстановленное изображение.

Как показала практика, данная задача имеет не единственное решение. Даже при малых искажениях изображения можно найти несколько решений в силу его неустойчивости. Поэтому необходимо добавить некоторую априорную информацию об исходном изображении. В этом случае задача восстановления изображения может быть сведена к задаче условной или безусловной оптимизации, в частности к задаче минимизации по Тихонову:

$$\hat{z} = \inf_f (\rho(Af, z) + \lambda\Phi(f)),$$

Функция $\Phi(f)$ должна гарантировать устойчивость и единственность решения. Ее вид определяется типом задачи. Она часто задаётся в виде L_2 нормы. Параметр регуляризации λ позволяет контролировать визуальное качество восстанавливаемого изображения и обычно также определяется на практике.

Данное уравнение можно переписать в рамках статистического подхода, используя максимальную апостериорную оценку (MAP):

$$\hat{x} = \arg\max_{x} \log p(y \mid x) + \log p(x),$$

где $p(y \mid x)$ – условная плотность вероятности, а $p(x)$ представляет собой априорную информацию об истинном изображении и не зависит шума.

С учетом этого, подход к обучению нейронной сети основан на минимизации заданной функции потерь для нахождения оптимального параметра Θ, характеризующего весовые коэффициенты нейронной сети. При этом большинство используемых функций потерь могут быть получены на основе MAP. В рамках нейросетевого подхода уравнение можно скорректировать следующим образом:

$$\hat{z} = \arg\min_{f} \rho(Af, z) + \lambda\Phi(f, \Theta),$$

где Θ – набор весовых коэффициентов нейронной сети

В настоящее время в задачах обработки изображений используется два больших класса ГНС: сверточные сети и сети-трансформеры, а также их возможные гибридные реализации. При этом можно представить обобщенную архитектуру таких сетей с учетом специфики решения задачи восстановления изображений в виде, показанном на рисунке 1.1.

В приведенной ниже схеме используется архитектура «encoder-decoder», в качестве составных частей которой будут как модули ГНС как сверточного, так и трансформерного типа. В конце процесса обработки происходит сложение исходного изображение с выделенными низкоуровневыми признаками от декодера, что является специфической особенностью задачи ВИ. В этом плане во многих известных работах было показано, что задача выделения шумов и искажений значительно легче, чем генерация полноценного восстановленного (улучшенного) изображения. Отдельно могут быть добавлены остаточные связи между каждым слоем кодера и декодера, как сделано, например, в Unet-подобных архитектурах.

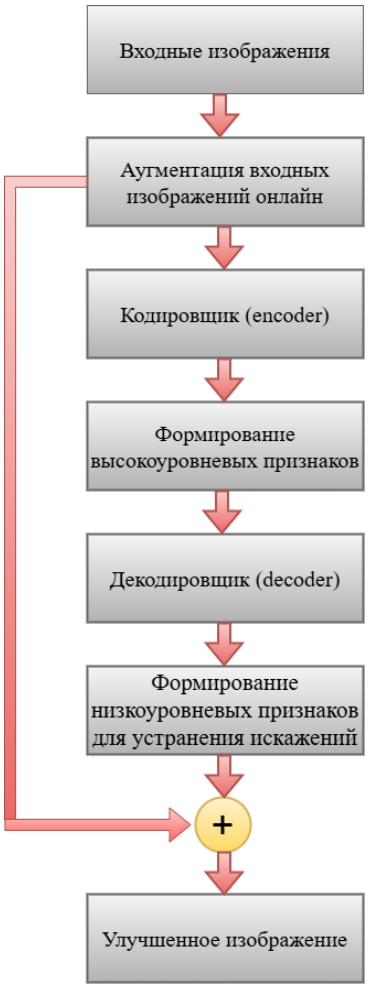


Рисунок 1.1. – Общая схема восстановления изображений

ГНС сверточного типа. Сверточные нейронные сети построены на использовании в кодировщике рис.1.1 последовательности сверточных слоев. Каждый такой слой реализует линейное преобразовании свертки, которое можно записать в следующем виде:

$$y_{i,j} = \sum_{-d<=a,b<=d} W_{a,b} x_{i+a,j+b},$$

где $y_{i,j}$ – результат свертки, x – входное изображение, $W_{a,b}$ – матрица обучаемых весов, размера $(2d+1)\times(2d+1)$.

После свертки необходимо добавить нелинейное преобразование, иначе сверточная нейронная сеть вырождается в простой линейный фильтр. Обычно для этого используют функцию $\mathrm{Re}LU$, представляющую собой функцию вида $\mathrm{Re}LU(x) = \max(x,0)$.

С каждым годом исследователи увеличивают количество сверточных слоев, повышая обобщающую способность нейронной сети. Первые слои используются для выделения обобщенных (низкоуровневых) признаков, в то время как последние слои служат для выделения уже более конкретных высокоуровневых признаков. В сверточных нейронных сетях также применяют механизмы внимания [12], которые определяют какие выходы нейронной сети в данный момент времени нужно учитывать больше всего, понимая локальный и глобальный контекст изображения.

Одной из типовых нейросетевых архитектур для восстановления изображений является архитектура «encoder-decoder». Основная цель ее состоит в отображении исходного изображения в скрытое пространство признаков с последующим его восстановлением [13] на основе признаков скрытого пространства. На рисунке 1.1 как раз и представлена обобщенная схема данной архитектуры.

Поворотным моментом для ГНС стала архитектура Inception (GoogLeNet, 2014) [14], предложившая идею параллельных многомасштабных сверточных блоков. В стандартном Inception-модуле объединяются свертки разного размера (1×1, 3×3, 5×5 и т.д.) и пулинг, что позволяет захватывать признаки различных масштабов одновременно. Такая многомасштабная обработка увеличивает выразительность сети без чрезмерного роста вычислительной сложности благодаря использованию 1×1 сверток для сокращения размерности каналов.

Архитектура Xception (2017) [15] пошла дальше, интерпретируя Inception модуль как частный случай разделимой свертки по глубине (depthwise separable convolution). Франсуа Шолле показал, что можно заменить сложные Inception-модули последовательностью из depthwise-сверток и последующей pointwise 1×1-свертки. Модель Xception («Extreme Inception») полностью строится на глубинно-разделимых свертках, достигая качества выше Inception V3 при том же числе параметров за счет более эффективного использования признаков.

Увеличение глубины сверточных нейросетей в начале 2010-х уперлось в проблему затухания градиентов и деградации качества при добавлении большого

количества слоев. Решением проблемы стала архитектура Residual Network (ResNet, 2015) [16] от Microsoft Research, которая ввела остаточные связи (residual connections, или skip connections). Идея ResNet – обучать не сами преобразования, а их приращения (residuals) так, чтобы выход каждого блока из сверточного слоя складывался с входом через остаточную связь.

Авторы ResNeXt (2017) [17], ввели понятие кардинальности – использование параллельных групповых сверток внутри блоков. В ResNeXt блоки выполняют несколько параллельных сверток меньшей размерности (grouped convolution) и объединяют результаты, что повышает выразительность без роста сложности. Кардинальность рассматривается как третий независимый гиперпараметр архитектуры наряду с глубиной и шириной.

Остаточные связи применяются в U-образных сетях (UNet) для сегментации и восстановления изображений – они помогают передавать низкоуровневые детали от входа к выходу, улучшая качество восстанавливаемых изображений. Даже при очень глубокой архитектуре наличие skip-связей гарантирует, что модель как минимум научится тождественному преобразованию и не ухудшит исходное изображение. Таким образом, ResNet не только позволил построить рекордно глубокие классификаторы, но и предоставил общий основу (backbone) для многих современных моделей компьютерного зрения.

Следующим шагом в развитии архитектур ГНС стала идея максимального повторного использования уже извлеченных признаков. Авторы DenseNet [18] (2016) предложили плотные соединения, где каждый слой получает на вход конкатенацию выходов всех предыдущих слоев в блоке. Иными словами, в плотном блоке существуют прямые связи от каждого слоя ко всем последующим. Это уменьшает затухание градиентов (как и ResNet), но дополнительно приводит к тому, что признаки низкого уровня передаются явно на все более высокие уровни и используются повторно. DenseNet удалось достичь сопоставимого с ResNet качества на ImageNet, но с существенно меньшим числом параметров, так как каждый слой добавляет лишь несколько новых карт признаков, полагаясь на объединение с ранее вычисленными признаками.

Еще один шаг в повышении эффективности ГНС – автоматический поиск архитектур (NAS, Neural Architecture Search) и масштабирование моделей. Инженеры Google в 2019 году представили EfficientNet [19], объединив оба подхода. Во-первых, базовая структура EfficientNet-B0 была найдена с помощью NAS, оптимизирующего точность при фиксированных вычислительных ограничения. Полученная архитектура применила инвертированные блоки MobileNetV2 с depthwise-свертками и остаточными связями. Во-вторых, EfficientNet ввела составное масштабирование (compound scaling): ранее масштабирование сетей происходило разрозненно – либо брали больше слоев (глубина), либо больше фильтров в каждом слое (ширина), либо более высокое разрешение входа. Авторы EfficientNet предложили совместно увеличивать эти параметры по заданным коэффициентам, сохраняя сбалансированность архитектуры.

В виду роста популярности трансформеров, авторы ConvNeXt [20] привнесли в ГНС сверточного типа ряд ключевых модификаций, влияющих на качество: отказались от небольших 3×3 фильтров в пользу более крупных (7×7 depthwise-свертки в блоках, наподобие размера окон внимания в Swin Transformer), заменили BatchNorm на Layer Normalization в определенных местах, убрали ReLU сразу после сверток, увеличили ширину блоков и размер внутренних MLP слоев (как в трансформерах). Также применили современные приемы обучения (например, адаптивные оптимизаторы, аугментация данных, регуляризация), которые стали стандартом для моделей-трансформеров. На самой большой модели ConvNeXt авторы достигли точности 87.8% Top-1 на ImageNet (224px) – уровень лучших Vision Transformer, а в задачах детекции (COCO) и сегментации (ADE20K) ConvNeXt даже превзошел Swin Transformer при сравнимых размерах модели.

Помимо этого, для восстановления изображений часто используются автокодировщиками. Основной смысл, которых состоит в создании тривиальной функции восстановления изображений [13]. На вход нейронной сети подается зашумленное изображение, по которому нейронная сеть пытается восстановить

исходное. В 2017 году авторами работы [21] была предложена модель RED-Net с остаточными связями, благодаря которым теряется меньше информации об исходном изображении, что увеличивает качество его последующего восстановления.

Проблемы восстановления изображения также исследуются на основе технологий обучения без учителя. В работе [22] авторы описывают Unet-подобную архитектуру, обучающуюся только на зашумленных изображениях с использованием техники «слепых пятен».

Описанные выше принципы авторы используют и при создании новых архитектур для восстановления изображений. Например, авторы IFSR-Net [23] (2025) предлагают архитектуру сверточной нейронной сети с неявным преобразованием частот, что позволяет эффективно восстанавливать изображения без использования преобразований Фурье. Модель демонстрирует улучшенные результаты по сравнению с традиционными методами восстановления изображений.

CV-CAN и CV-DDAN (2024) [24] – модели, основанные на комплекснозначных сверточных нейронных сетях (CV-CNN), которые используют механизмы внимания на комплексных числах для восстановления изображений в частотной области. Показали превосходство над реальными аналогами в задачах удаления шума и повышения разрешения изображений.

Авторы VmambaIR (2024) [25] используют визуальную модель состояния (Visual State Space Model) для восстановления изображений, которая использует механизмы Omni Selective Scan (OSS) и Efficient Feed-Forward Network (EFFN) для эффективного моделирования информации в изображениях. Модель достигает передовых результатов в различных задачах восстановления изображений при меньших вычислительных ресурсах.

Исследователи в KBNet (2023) [26] используют модуль внимания на основе ядра (Kernel Basis Attention), который вводит обучаемые базисные ядра для моделирования представительных образов изображений. Модель достигает

передовых результатов в задачах восстановления изображений, удаления плохих погодных осадков при меньших вычислительных затратах.

В современных реалиях становится нормой, что новая архитектура заявляет не только улучшение качества, но и лучшее соотношение точность/расходы. В контексте задач вроде восстановления изображений это особенно важно, когда модели должны работать, например, на камерах смартфонов или в реальном времени на видео. Поэтому архитектуры для суперразрешения, восстановления изображений от шума и пр. заимствуют или вводят новые блоки для увеличения эффективности работы нейронной сети. Тенденция такова: делать модели легче, быстрее, но при этом умнее за счет продуманных архитектурных решений, а не за счет аппаратного роста и увеличения количества слоев.

ГНС трансформерного типа. В последнее время для решения различных задач в области компьютерного зрения все большую популярность набирает использование моделей-трансформеров (vision transformer, ViT), впервые описанных в работе [27]. Данная архитектура появилась в области обработки естественного языка (NLP), где показала свою эффективность с помощью реализации механизма внимания (самовнимания) [12]. В настоящее время проводятся многочисленные исследования, направленные на сравнение трансформеров со сверточными нейронными сетями (СНС). Показано, что механизм внимания, реализуемый в ViT, в функциональном плане во многом эквивалентен операции свертки [28]. При этом по эффективности модели-трансформеры во многих случаях могут давать лучшие результаты по сравнению со сверточными сетями с гораздо большим количеством слоев в задачах классификации, семантической сегментации, повышения качества изображений.

Одним из важных мест в моделях трансформеров, используемых для улучшения изображений, как уже упоминалось, является реализация механизма внимания. Основная цель его состоит в том, чтобы выбрать наиболее значимые признаки из исходных изображений, на которых нужно сосредоточиться и которые наиболее важны в ходе последующей обработки для повышения эффективности улучшения изображений. В большинстве случаев используется

его разновидность – самовнимание, когда все значения берутся непосредственно с выхода одного слоя нейронной сети, путем умножения на соответствующие матрицы коэффициентов [29].

В первоначальной архитектуре ViT изображение разбивается на блоки 16x16. Из-за того, что модели-трансформеры инвариантны к перестановкам пикселей, авторы [27] добавляют еще позиционное кодирование для понимания взаимного расположения блоков. Далее используется стандартный для NLP блок многоголового внимания, на выходе которого применяются полносвязные слои для решения задачи классификации.

Исследователями было выявлено, что данная архитектура показывает невысокие результаты в задачах улучшения качества изображений, сегментации и детектирования объектов. Одна из основных проблем – разделение на блоки 16x16 пикселей, что приводит к снижению точности результатов. Однако, уменьшение размерности блоков нежелательно, так как это существенно увеличивает вычислительную сложность нейронной сети.

Авторы работы [30] предлагают собственный блок swin-трансформер (shifted windows transformer) для реализации локального механизма внимания с использованием сканирующего окна, а также добавлением обучаемых параметров для кодирования блоков изображения. При этом уменьшается размер блоков до 4x4 пикселей, а также добавляется иерархичное их объединение. Это позволяет сделать постоянным число признаков на разных масштабах, поступающих на вход механизма внимания. Данная архитектура уже успешно использовалась в задачах обнаружения объектов и семантической сегментации [31-32].

В работе [33] предлагается усовершенствованный механизм внимания трансформеров для задачи семантической сегментации изображений. Авторы решают уменьшить вычислительную сложность блока внимания путём сокращения размерности изображения в R раз с помощью так называемой разделяемой по глубине (depthwise) свертки с таким же шагом R.

В работе [34] swin-трансформеры используются в задаче устранения шумов и искажений на изображении. Здесь авторы предлагают архитектуру SUNet,

построенную на подобии и принципах архитектуры UNet. Они оставляют только сверточные слои на входе и на выходе, а также блоки повышения и понижения дискретизации. Все остальные блоки представляют собой swin-трансформеры.

В контексте рассматриваемой задачи возможно использование моделей-трансформеров в генеративно-состязательных сетях. Авторы [35] при помощи двух дискриминаторов с остаточными связями от генератора улучшают качество изображений при низком освещении. Помимо этого, на вход нейронной сети добавляется градиент изображения, как априорная информация о его структуре. Это позволяет успешно восстанавливать текстурные части изображения.

В работе [36] авторы для повышения разрешения изображений применяют блок MDTA (multi-Dconv head transposed attention), который использует межканальную ковариацию для получения оптимальных карт внимания. Преимущество MDTA заключается в том, что он анализирует глобальные взаимосвязи между пикселями изображения и оптимизирует локальный контекст для выделения признаков с целью последующей обработки в плане улучшения качества изображений. Помимо этого, авторы используют на выходе из MDTA depthwise свертки с вентильным механизмом на основе функции активации GELU вместо общепринятых полносвязных слоев.

В работе [37] авторы предлагают архитектуру Transweather – модель-трансформер, ориентированная на восстановление изображений в плохих погодных условиях. Авторы используют одну модель и обучают нейронную сеть одновременно для различных видов погодных осадков.

Авторы большой работы [38] борются с проблемой нехватки данных через самообучение (Self-Supervised Learning), когда важно не столько решение самой задачи повышения качества изображений, а сколько выделение признаков, которые будут получены в ходе ее решения. Выделенные признаки можно в дальнейшем использовать при обучении нейронной сети в задачах с маленьким набором размеченных данных.

Помимо этого, многие исследователи используют технику переноса обучения [39], когда нейронная сети обучается в два этапа: сначала на большом

размеченном датасете, а потом уже на относительно небольшом, ориентированным на конкретную задачу. Также применяются различные методы аугментации данных.

Описанные выше подходы к обучению нейронных сетей используются и в данной работе. При этом при подготовке обучающих данных в работе применяются собственные алгоритмы аугментации данных [40], позволяющие генерировать различные виды шумов и моделировать реальные искажения.

Изначально в работе [41] выделено несколько основных проблем при использовании трансформеров для решения задач компьютерного зрения:

1. Длительный процесс обучения и сложная интерпретируемость результатов.
2. Квадратичная вычислительная сложность относительно числа пикселей изображения из-за использования механизмов внимания.
3. Необходимость использования большого набора данных для обучения.

На сегодняшний день эти проблемы остаются актуальными. Описанные в известных работах подходы отчасти решают проблему производительности и получения требуемого количества обучающих примеров, но не всегда позволяют использовать трансформеры в реальном времени на небольшом наборе данных в задачах компьютерного зрения.

1.2. Алгоритмы аугментации данных при решении задач восстановления и улучшения качества изображений

Нейронные сети очень чувствительны к набору данных и требуют большого их количества при обучении. Однако существующих датасетов в многих случаях недостаточно для решения современных задач машинного обучения. Большинство из них создано для обучающих целей и мало пригодно для решения конкретных практических задач. Поэтому исследователи вынуждены проводить дополнительную разметку обучающих данных, что может быть затратно во всех отношениях. Кроме того, в некоторых случаях получить новые данные технически сложно и дорого, например, данные по сейсмической активности или

изображений сцен в условиях атмосферных осадков. При этом важно, чтобы распределение генерируемых данных было максимально похоже на искомую генеральную совокупность изображений. Чем точнее это приближение, тем выше будет качество работы нейронной сети. Другим примером является задача восстановления изображений при наличии дефектов и другого рода искажений, вызванных различного рода артефактами. Существует большое количество типов таких искажений и, в то же время, недостаточно реальных данных для обучения, так как дефекты производства, фиксируемых изображениями, возникают достаточно редко. Поэтому задача генерации реалистичных дефектов остается актуальной на сей день.

Аугментация данных может применяться для любого вида и формата данных. Существует множество алгоритмов для работы с графической, звуковой и текстовой информацией. При этом для каждого из видов информации необходим свой вариант преобразования. Например, для аугментации текстовых данных исследователи [42, 43] используют метод обратного перевода, что может существенно увеличить количество текстовых данных, однако, для других видов информации это совершенно бесполезно.

Понимая разнообразие и сложность графических данных в дальнейшем ограничимся только алгоритмами аугментации, применяемые по отношению только к изображениям.

Различают два принципа аугментации: онлайн и офлайн. Офлайн подход заключается в непосредственном добавлении аугментированных данных ко всему датасету и использование их во время обучения. Онлайн подход или аугментация на лету трансформирует данные на каждой эпохе обучения нейронной сети, искаженные результаты никуда не сохраняются. В современных исследованиях по большей части используется второй подход, так как он не требует дополнительного места на жестком диске и позволяет создавать большее количество разнообразных данных. Одним из первых применений онлайн-аугментации стало обучение сети AlexNet, которая заняла первое место по точности на датасете ImageNet в 2012 году [44].

В независимости от подхода и метода аугментации необходимо частично сохранять исходное распределение данных. Следовательно, каждое из видов искажений должно применяться с какой-то долей вероятности. Помимо этого, если возникают схожие искажения, то необходимо создавать политики аугментации, а далее на каждой итерации к изображению применять одну из этих политик [45].

Аугментация позволяет увеличить количество данных, повысить скорость разработки и уменьшить ее стоимость. Более того, она может способствовать уменьшению количества параметров нейронной сети, так как нейронная сеть начинает выделять общие признаки, а не специфичные для конкретного датасета.

Одним из недостатков этого подхода является слабая связь между качеством работы нейронной сети на тестовых и реальных данных. Помимо этого, не всегда достаточно применения простых техник аугментации: некоторые задачи требуют специфичных изображений. Поэтому в этих случаях необходимо использовать более продвинутые техники аугментации, не только реалистично искажающие исходные изображения, но и генерирующие и синтезирующие новые изображения.

Существует огромное множество алгоритмов аугментации, создающие близкие к реалистичным искусственные обучающие данные. Общая их классификация представлена на рисунке 1.2 [46].

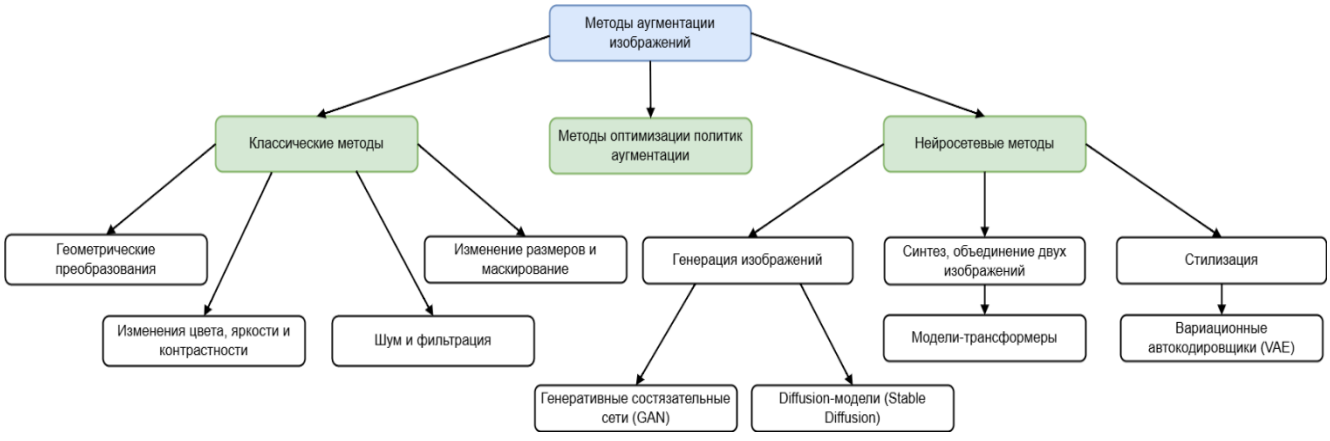


Рисунок 1.2 – Классификация алгоритмов аугментации данных

Классические методы аугментации изображений включают различные геометрические и фотометрические преобразования, направленные на увеличение разнообразия обучающих данных и улучшение обобщающей способности моделей машинного обучения. Эти методы позволяют моделям обучаться на различных вариантах представления изображений, повышая их устойчивость к изменениям в данных. Добавление шума и частотная фильтрация также используются для повышения устойчивости моделей к различным искажениям и шумам в данных.

Классические алгоритмы сталкиваются с проблемами разнообразия изображений и, часто, недостаточной реалистичности аугментированных данных [47]. Нейросетевые алгоритмы позволяют решить эти проблемы, при этом методы глубокого обучения самостоятельно подбирают оптимальные параметры для аугментации. В последнее время набирают популярность методы, основанные на генеративно-состязательных сетях (GAN) и диффузионных моделях, а также трансформерные модели для генерации данных из-за возможностей механизма внимания улавливать контекст и важные детали исходного изображения. В настоящей работе автором как раз показывается эффективность применения последнего подхода.

Отдельным направлением развиваются политики аугментации данных, решая проблему разнообразия сгенерированных данных в первую очередь для классических алгоритмов. Политики аугментации [48-49] представляют собой стратегии выбора и применения различных методов аугментации для оптимизации обучения моделей. Например, AutoAugment [48] использует методы обучения с подкреплением для автоматического поиска оптимальных комбинаций аугментаций, специфичных для конкретного датасета. Этот подход позволяет существенно улучшить точность моделей, однако требует значительных вычислительных ресурсов для обучения. RandAugment [49] упрощает процесс аугментации, случайным образом выбирая фиксированное количество аугментаций с заданной интенсивностью. Этот метод обеспечивает сопоставимую

с AutoAugment производительность при значительно меньших вычислительных затратах.

Описанные выше алгоритмы также можно разбить на два больших класса: порождающие и дискриминантные. Дискриминантные модели отображают вход x в выход y и с математической точки зрения реализуют условное распределение $p(y|x)$. В отличии от них порождающие генеративные алгоритмы ставят задачу формирования фактически $p(x|y)$ или же $p(x)$, если необходимо генерировать случайные элементы из всей генеральной совокупности.

В последнее время все большую популярность набирают порождающие алгоритмы в виду их реалистичной генерации и обобщающей способности [50]. Математическую задачу оптимизации для порождающих моделей можно сформулировать на основе метода максимального правдоподобия следующим образом:

$$\theta' = \arg\max_{\theta} \sum_{i=1}^{N} \log p(x_i;\theta),$$

где θ – параметры модели; x_i – элемент из набора обучающих данных.

1.2.1. Эвристические алгоритмы преобразования изображений

В таких алгоритмах ставится акцент на аугментацию изображений путем добавления дополнительных шумов и искажений. Пусть функция $I(x,y)$ описывает исходное изображение, а $J(x,y)$ – является его преобразованным изображением. Можно выделить следующие типы искажений [51].

1. Геометрические искажения (аффинные, проективные):

$$J(x,y) = I(f_1(x,y), f_2(x,y)),$$

где f_1 и f_2 функции, определяющие геометрические преобразования. В случае поворота они будут выглядеть так:

$$f_1(x,y) = x\cos\psi + y\sin\psi, \quad f_2(x,y) = -x\sin\psi + y\cos\psi,$$

где ψ определяет угол поворота.

Изменение масштаба в рамках этой модели можно задать следующим образом:

$$J(x,y) = I(e^{\psi_1}x, e^{\psi_2}y),$$

где ψ_1, ψ_2 определяют соответствующие параметры изменения масштаба.

2. Глобальное изменение яркости и контраста

$$J(x,y) = f(I(x,y)),$$

где f – искажающая функция, на которую накладываются ограничения независимого изменения пикселей. Следовательно, пиксели с одинаковым значением яркости должны сохранить одинаковую яркость после преобразования. Очень часто накладывают дополнительные ограничения на данную функцию, например, монотонность или биективность. В частном случае уравнение можно записать следующим образом:

$$J(x,y) = \psi_1 + \psi_2 I(x,y).$$

3. Локальное изменение яркости

$$J(x,y) = f(\{I(x',y'),(x',y') \in O(x,y)\}),$$

где $O(x,y)$ определяет окрестность точки (x,y).

4. Зашумления и искажения, характерные для решаемой задачи: блики, шумы, размытие и др. Их можно описать следующей формулой:

$$J(x,y) = f(I(x,y), \xi_{x,y}),$$

где $\xi_{x,y}$ независимая случайная величина, обычно использующаяся для моделирования аддитивного белого гауссовского шума.

5.Аппликация. В частном случае она может использоваться для моделирования аппликативных и импульсных помех. Преобразование задаётся следующим образом:

$$J(x,y) = M(x,y)I_1(x,y) + (1 - M(x,y))I_2(x,y),$$

где $M(x,y)$ – изображение маска и $0 \le M(x,y) \le 1$; I_1 и I_2 – изображения, которые накладываются.

Данное преобразование можно представить, как общий случай стилизации и синтеза изображений. Основной его недостаток в том, что параметры маски под каждое изображение необходимо подбирать экспериментально.

Для данных видов искажений параметры ψ обычно настраиваются экспериментальным путем. В тоже время аугментация может быть чувствительна к подбору параметров алгоритма искажения. В большинстве случаев они задаются перед началом обучения. Возможны другие подходы, например, это можно сделать с помощью проверки качества и нахождения трудных результатов на проверочной выборке или же использовать библиотеку Google AutoAugment [48], которая автоматически подбирает параметры для обучения.

1.2.2. Генерация изображений с помощью глубоких нейронных сетей

На сегодняшний день не существует универсального алгоритма генерации всевозможных типов помех на основе ГНС. Исследователи сосредотачиваются на моделировании их отдельных типов.

Отдельно стоит отметить вариационный автокодировщик (VAE). Автокодировщики позволяют эффективно извлекать признаки из изображений, обучаясь на тривиальной задаче их восстановления. Однако пространство признаков оказывается достаточно разреженным, что является серьезной проблемой, поскольку найти скрытое значение, для которого декодер будет знать, как сгенерировать нормальное изображение, почти невозможно. На помощь приходит VAE, он позволяет сделать скрытое пространство непрерывным и оценить плотность распределения обучающих данных. Это происходит за счет явной параметризации исходного неизвестного распределения путем задания математического ожидания и стандартного отклонения в качестве обучаемых параметров [52].

Для генерации изображений может также использоваться нейронная сеть PixelRNN и ее разновидности. Суть идеи в использовании рекуррентных нейронных сетей для предсказания пикселя изображений на основе уже

известных [53]. Оценку совместного распределения пикселей можно записать в следующем виде: $p(X) = \prod_{i=1}^{N^2} p(x_i \mid x_1, x_2, ...x_{i-1})$. Подход позволяет получить изображения среднего качества, при этом обладает значительным преимуществом – быстрая скорость обучения нейронной сети.

Описанные выше порождающие алгоритмы в большинстве случаев не могут служить для аугментации данных, так как некоторые из них существенно искажают исходное распределение, генерируя не вполне реалистичные данные, а другие имеют проблемы в производительности, что не применимо для аугментации данных на лету.

На сегодняшний день большого успеха и распространения в генерации изображений добились генеративно - состязательные сети. Они позволяют формировать принципиально новые изображения на основе уже существующих, однако имеют большую вычислительную сложность по сравнению с алгоритмами стилизации изображений. GAN в большинстве случаев не требуют парных данных для обучения, при этом они могут обучаться без учителя. В статье [54] приводится пример использования простой архитектуры GAN для генерации реалистичных царапин на металле, возникших в процессе производства.

GAN состоит из двух подсетей: генератора и дискриминатора. Одна сеть пытается научиться порождать правильные примеры, обманывая вторую, а вторая – отличить сгенерированные картинки от настоящих. По мере обучения они постепенно делают друг друга лучше.

Формально генератор можно задать следующим образом:

$$G = G(z, \theta_g) : Z \to X ,$$

где Z – некоторое пространство скрытых факторов, имеющее априорное распределение.

Тогда дискриминатор задаётся таким образом:

$$D = D(x, \theta_d) : X \to [0,1] .$$

Определим функцию оптимизации для GAN как:

$$\max_D V(D) = E_{x\sim p_{data}(x)} \log D(x) + E_{z\sim p_z(z)} \log(1 - D(G(z))),$$

$$\min_G V(G) = E_{z\sim p_z(z)} \log(1 - D(G(z))),$$

где E обозначает математическое ожидание по всему распределению данных. Первый член в формуле показывает, насколько хорошо дискриминатор определяет реальные данные, а второй член формулы отвечает за то, на сколько хорошо генератор «обманывает» дискриминатор.

Представленные выше формулы можно объединить в одну:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)} \log D(x) + E_{z\sim p_z(z)} \log(1 - D(G(z)))$$

Данная формула отражает всю суть GAN. В последнее время исследователи добавляют дополнительные слагаемые в формулу, увеличивая устойчивость нейронной сети. Также для увеличения точности работы и генерации изображений в хорошем качестве исследователям приходиться значительно углублять архитектуру нейронной сети на примере BigBiGAN [55]. Однако ее эффективность в режиме реального времени ставится под сомнение.

В качестве простых архитектур GAN авторы [56] используют SDGAN – продвинутую версию CycleGAN для решения задачи преобразования изображений без дефектов в изображения с дефектами. В тоже время у данной работы есть существенный недостаток. Необходим однородный датасет, содержащий изображения как с дефектами, так и без них. Другим минусом данной работы является то, что для каждого нового типа дефектов необходимо заново обучать нейронную сеть.

Существующее большое множество архитектур GAN и способов генерации ими изображений в большинстве случаев их можно разделить на две категории: условная и безусловная генерация. В первом случае к шуму добавляются метки классов и используются Embedding слои. Во втором случае подается только случайный гауссовский шум и данную задачу можно рассматривать как обучение без учителя, так как не требуются дополнительные метки.

В последнее время становится популярной идея использовать усеченное нормальное распределение при генерации тестовых данных. Основная идея

состоит в том, чтобы уменьшить разброс значений при генерации шума. Нейронная сеть, получая незнакомый вход, который не был ей известен во время обучения, генерирует изображения плохого качества. Следовательно, если уменьшить разброс, то можно найти компромисс между разнообразием и качеством самих изображений.

Несмотря на успехи в применении GAN, обучать их достаточно трудно. Во-первых, это решение сложной оптимизационной задачи maxmin и нахождение седловой точки. Во-вторых, часто возникает неустойчивость при обучении. В статье [57] описываются основные популярные архитектуры GAN и теоретически доказывается их устойчивость с помощью теории векторных полей. В работе не представлены подробные математические выкладки, но всегда можно математически обосновать выбор каждой функции потерь.

Другая проблема заключается в том, что в большинстве случаев задача дискриминатора значительно легче задачи генератора. Дискриминатор учится быстрее и в скором времени начинает определять все результаты генератора как ложные. Из-за этого генератор не знает, в какую сторону улучшить результат генерации. Помогает решить эту проблему использование различных функций потерь вместо бинарной кроссэнтропии и особая регуляризация для дискриминатора (WGAN). В CycleGAN используется Patch дискриминатор, на его выходе формируется не 0 или 1, а целая карта признаков, исходя из которой, генератор может точнее понять, где он делает ошибку.

Помимо этого, может возникнуть неприятная ситуация, когда генератор генерирует небольшое количество разнообразных изображений хорошего качества, тем самым обманывая дискриминатор. Однако практическая ценность его невысока. Данная проблема в англоязычной литературе фигурирует, как «mode collapse». Однозначного способа ее решения сейчас не существует, но исследователи вводят различные метрики для оценки качества модели и выявления этой проблемы.

Другой негативной стороной GAN является большое число параметров сети, а также необходимость их тонкой настройки. Авторы статьи [58] исследуют

различные архитектуры GAN при условной генерации изображений и приходят к выводу, что при дефолтном назначении всех параметров различные архитектуры показывают практически одинаковый результат. Но, проведя их оптимизации путем изменения скорости обучения, количества фильтров в сверточных слоях, количества узлов в полносвязном слое и др., можно существенно увеличить эффективность GAN. Это доказывает, что авторы статей настраивают параметры нейронной сети под определенные датасеты и в общем случае она теряет обобщающую способность.

Исходя из этого, начинают набирать популярность диффузионные модели. Основная идея которых заключается в последовательном добавлении к исходному изображению шума, с дальнейшим его восстановлением. Данный подход хорошо показал себя как при условной, так и безусловной генерации изображений, а также в задаче сверх разрешения. В одной из последних работ [59] авторы путем использования глубокой архитектуры сверточной нейронной сети и дополнительных механизмов внимания превзошли по точности современные модели GAN и на сегодняшний день являются одними из лучших в генерации изображений из датасета ImageNet.

1.2.3. Синтез изображений реальных сцен в условиях атмосферных осадков

Данная задача особенно актуальна, поскольку во многих практических ситуациях требуется формировать изображения реальной сцены для обучения нейронных сетей с целью повышения качества регистрируемых изображений (например, с помощью видеокамеры) и понимания сцены в условиях различных атмосферных осадков (плохих погодных условиях). Подобных данных для адекватной разметки и включения их в процесс обучения всегда не хватает, что определяет важность применения методов аугментации для синтеза изображений одних и тех же объектов в различных погодных условиях.

В этом плане следует опять отметить GAN, которые научились переносить стили между изображениями. В контексте поставленной задачи модели GAN

можно обучить превращать «чистое» изображение в «дождливое», например, pix2pix GAN [60]. Но в реальности собрать идеально совпадающие пары для обучения крайне трудно – нужно сфотографировать одну и ту же сцену при различных атмосферных осадках и без них. Поэтому исследователи используют «непарные» методы перевода изображений как, например, в модели CycleGAN [61]. В данной архитектуре вводится цикл последовательных переводов изображения (чистое→дождливое→чистое) с требованием восстановить исходный снимок. Модель CycleGAN и производные от нее модели успешно выполняют перенос изображений погоды между несвязанными наборами данных. Однако, отсутствие прямого контроля качества обучения в указанной архитектуре приводит к непреднамеренным искажениям: могут появляться несуществующие объекты или артефакты на изображениях. Последующие исследования вводят ограничения для сохранения содержания и структуры изображения. Например, модели MUNIT и DRIT [62-63] разделяют скрытое представление на контент и стиль, что позволяет выполнять мультимодальное отображение. Это даёт возможность сгенерировать несколько вариантов «дождливой» сцены из одной исходной, меняя только вектор стиля. Несмотря на сохранение структуры изображения, у таких моделей все еще остаются проблемы с реалистичностью и геометрией осадков.

В последнее время стали популярны диффузионные модели, изначально успешные в генерации изображений по тексту. Диффузионные модели способны более устойчиво обучаться и лучше моделируют сложные детали, чем GAN [64-65]. Однако при прямом их использовании для добавления атмосферных осадков на изображения исследователи сталкиваются с проблемами производительности и скорости генерации. В тоже время, в 2024 году была предложена нейронная сеть CycleGAN-Turbo (Parmar et al.) [66], объединяющая компоненты диффузионной модели и GAN в единый генератор, способный сохранять структуру изображения и обеспечивать высокое качество переноса стиля. Модель обучается без парных данных и за один проход добавляет туман, снег или дождь; при этом она превосходит по качеству предыдущие GAN и диффузионные методы.

Помимо этого, исследователи используют уже обученную диффузионную модель StableDiffusion и Condition/ControlNet механизмы [67]. Так в работе (Greenberg et al., 2024) [68] авторы предлагают новый метод преобразования изображений Seed-to-Seed Translation (StS). Он использует пространство «инвертированных сидов» для перевода и обучает sts-GAN для трансформации между сид-состояниями без необходимости иметь набор парных данных. Авторы добились того, что можно преобразовать дневное чистое изображение в ночное или дождливое, сохраняя геометрию объектов на изображении, благодаря управлению через карты контуров ControlNet.

Отдельно стоит отметить мультимодальный подход для решения указанной задачи. Мультимодальность в данном случае понимается как использование текстовых описаний сцены наряду с изображением. Исследователи начинают объединять большие языковые модели и генеративные модели для генерации сложных сцен на изображениях. Например, в WeatherDG (Qian et al., 2024) используется большая языковая модель, которая генерирует подробный текстовый сценарий, а затем диффузионная модель Stable Diffusion, дообученная на исходных данных по этому описанию, рисует изображение в требуемых погодных условиях [69]. WeatherDG способна генерировать разнообразные дорожные сцены при дожде, тумане, снеге, причём добивается того, чтобы даже редко встречающиеся объекты (например, пешеходы, мотоциклы) корректно отображались при плохой погоде. Такие мультимодальные решения фактически создают не просто перенос стиля с конкретного изображения, а делают синтез новых.

В целом, современные исследования направлены на достижение максимально точного и управляемого наложения погодных эффектов, сохраняющих каждый элемент структуры исходного изображения. В итоге можно выявить следующие тенденции.

Комбинация методов: объединение физических моделей осадков с обучением нейронных сетей как в TRG-Net [70]. Генерируемые изображения, в

таком случае, выглядят более правдоподобно, при этом их качество может контролироваться пользователем.

Улучшение качества и реалистичности при аугментации данных: исследователи добиваются того, чтобы сгенерированные изображения со снегом и дождём были настолько реалистичны, что их можно использовать для тренировки и тестирования других систем (детекторов, сегментаторов) без потери качества. Это своего рода тест Тьюринга для «погодного синтеза» – можно ли обучить модель на сгенерированных данных и получить ту же или близкую эффективность, что и на реальных? Пока что полностью заменить реальный мир синтетическими данными не удалось, но с каждым годом разрыв быстро сокращается.

Универсальность и адаптация: создаются модели, способные одним набором весов генерировать разные виды погодных явлений и под разные входные изображения. Например, CycleGAN-Turbo одним генератором добавляет или убирает дождь и туман, а также меняет время суток. В то же время, общие многофункциональные модели WeatherDG являются мультимодальными.

Точный перенос: исследователи ставят задачи, которые требуют точного переноса конкретного шаблона осадков. Это достигается, когда модели, например, Mask-DerainGAN [71] могут выделить шум осадков как отдельный слой и перенести его на новое изображение. Однако диффузионные и особенно мультимодальные генераторы пока больше заточены под стилевой перенос. Поэтому узкоспециализированные методы с использованием масок погодных осадков, раздельной обработки фона все еще актуальны, особенно когда нужен идентичный по стилю дождь, снег, туман на сгенерированном изображении.

Несмотря на достигнутый прогресс, остаются проблемы с генерацией очень сложных и специфичных случаев, например, «ливень ночью со встречным светом фар», где капли образуют блики; добавление осадков на видео с учетом движения; перенос таких эффектов, как иней, лужи, снегопад с разной глубиной резкости. GAN-подходы, такие как CycleGAN, DRIT и MUNIT [60-62], хотя и способны сохранять структуру изображения, но часто приводят к искажениям геометрии

осадков и появлению несуществующих артефактов из-за отсутствия прямого контроля качества. Современные диффузионные модели, в том числе методы на основе Stable Diffusion с ControlNet, значительно улучшают реалистичность и управляемость переноса стиля и погодных осадков, но по-прежнему требуют значительных вычислительных ресурсов и остаются достаточно медленными, особенно при обработке изображений высокого разрешения. Мультимодальные методы (например, WeatherDG) показывают впечатляющие результаты, но требуют сложного многоступенчатого процесса обучения, интеграции языковых моделей и дообучения диффузионных моделей на больших специализированных датасетах. Это ограничивает их применение в практических задачах с малым набором данных и низкими ресурсными затратами.

В связи с изложенным обозначенная проблема остаётся актуальной по сей день и требует создания универсальных архитектур, позволяющих максимально точно и реалистично генерировать различные погодные осадки и искажения на реальных изображениях. Остаётся необходимость дальнейших исследований и разработки новых решений, способных одновременно обеспечить высокую точность, скорость генерации и низкую вычислительную сложность, что и явилось основной мотивацией для данной работы.

1.3. Постановка задачи и общая схема проведения исследований в интересах построения алгоритмов восстановления изображений и аугментации данных

С точки зрения организации процесса восстановления и синтеза изображений в условиях шумов и искажений, особый интерес представляет объединение методов улучшения качества изображений и методов генерации и синтеза данных в единую архитектуру, устойчивую к переобучению и способную обобщать и систематизировать знания в различных задачах компьютерного зрения. На современном этапе развития компьютерного зрения показано, что применение только одного из подходов – либо восстановление, либо аугментация – недостаточно для построения универсальных решений. Требуется комплексный

подход, включающий анализ, синтез, генерацию и обучение. В соответствии с этим автором была предложена общая схема проведения исследований, показанная на рис.1.3.

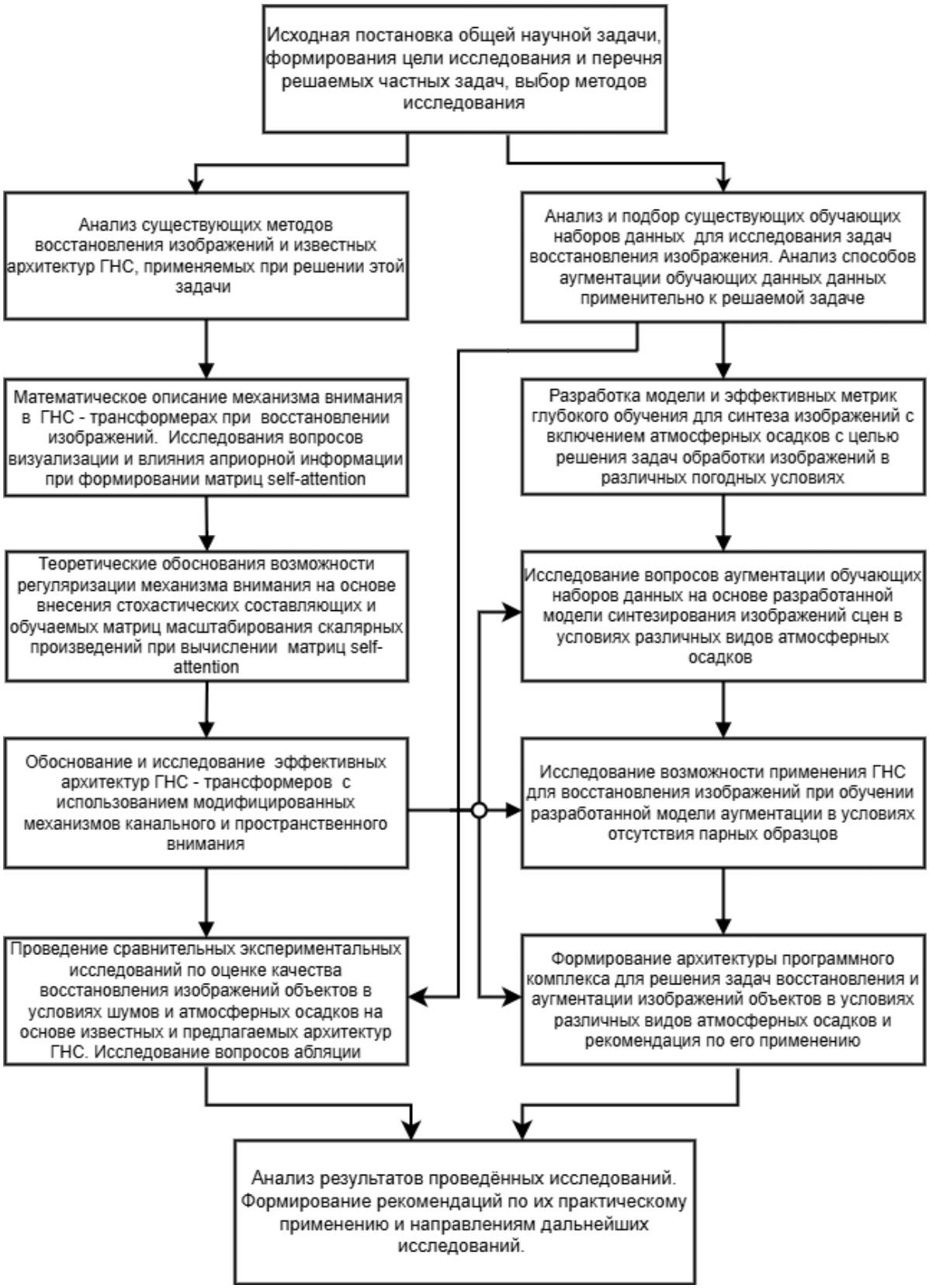


Рисунок 1.3 – Общая схема проведения исследований

В качестве отправной точки исследований выступает анализ существующих методов восстановления изображений (глава 1, п.п. 1.1.1-1.1.3), включая как классические подходы, так и современные архитектуры глубоких сверточных сетей и сетей трансформерного типа. Проведенный обзор выявил ограниченность классических алгоритмов в условиях сложных и пространственно неоднородных искажений, а для алгоритмов, основанных на использовании нейронных сетей, недостаточную устойчивость к шумам необходимость большого объема данных при обучении. Это послужило обоснованием для выделения двух ключевых направлений развития: исследование и модификации механизмов внимания в моделях-трансформерах в задачах ВИ и построение эффективных алгоритмов аугментации данных и, прежде всего, для задач восстановления.Для решения поставленной общей задачи в схеме выделены две параллельные взаимосвязанные ветви исследований, соответствующие двум типам задач обработки изображений: алгоритмы восстановления изображений с помощью моделей-трансформеров и алгоритмы аугментации данных для различных задач компьютерного зрения.

Первая ветвь (главы 2) направлена на теоретическое обоснование модификаций механизмов внимания, лежащих в основе моделей ГНС трансформерного типа. Рассматриваются методы регуляризации внимания путем добавления стохастических, мультипликативных и корреляционных составляющих (п. 2.3.1–2.3.3), обеспечивающих устойчивость и адаптивность модели к изменениям во входных данных путем сглаживания быстро растущих весов внимания. Вводится обучаемая матрица масштабных коэффициентов скалярных произведений матриц внимания (п. 2.3.4), помогающая бороться с эффектами насыщения функций активации. По результатам теоретических исследований были обоснован и синтезированы архитектуры трансформеров с модифицированными механизмами канального и пространственного внимания, ориентированные на задачу восстановления изображений (глава 3, п.п. 3.1–3.2). Проведено их экспериментальное исследование на синтетических и реальных искажениях (п. 3.3).

Вторая ветвь (глава 4) посвящена разработке алгоритмов аугментации данных, включая эвристические преобразования изображений, для которых установлены их ограниченные возможности с точки зрения способности реалистично генерировать помехи. Показываются преимущества глубоких нейронных сетей на основе собственной архитектуры модели-трансформера синтеза изображений в плохих погодных условиях с помощью с использованием изображений-шаблонов. Учитывая сложность и разнообразие реальных искажений, особое внимание уделено моделированию атмосферных осадков в виде дождя, снега, тумана и других погодных явлений, а также их реалистичному переносу на другие изображения (п. п. 4.2–4.3). Проведено сравнение различных методов аугментации и анализ их влияния на производительность нейросетей в задачах восстановления (п. 4.4).

Обе ветви исследований объединяются на этапе синтеза универсальной архитектуры восстановления и синтеза изображений, использующей как модифицированные трансформеры, так и адаптивные алгоритмы аугментации. Предложенная архитектура сочетает в себе трансформер с модифицированным механизмом внимания и процедуру синтеза зашумленных и искаженных изображений. Это позволяет решить сразу две задачи: повысить устойчивость модели в условиях сложных и редко встречаемых искажений и расширить доступную обучающую выборку без необходимости ее ручного сбора.

На завершающем этапе исследований (главы 3.4 и 4.5) проведено комплексное экспериментальное сравнение синтезированных архитектур с современными аналогами. Используются как стандартные метрики (PSNR, SSIM, FID), так и введенные в работе метрики, учитывающие аспекты визуального восприятия синтезируемых изображений и оценка работоспособности нейронных сетей на реальных данных. По результатам экспериментов сформулированы рекомендации для дальнейших исследований, включая направления по улучшению внимания, генерации сложных погодных условий, а также интеграции мультимодальных входов (например, текста и карты контуров) в архитектуру генераторов.

Таким образом, предложенная схема охватывает весь цикл исследований: от теоретического анализа до практической реализации и экспериментальной проверки. Она демонстрирует, что сочетание глубоких архитектур с адаптивной генерацией и синтезом данных, дает существенные преимущества в устойчивости, эффективности и точности систем компьютерного зрения.

Выводы по главе

1.Проведен анализ существующих подходов к задаче восстановления и аугментации изображений. Выявлены ограничения классических методов обработки, основанных на линейной и нелинейной фильтрации, частотных преобразованиях и регуляризации. Установлено, что данные методы обладают ограниченной универсальностью, высокой чувствительностью к выбору параметров и неэффективны при наличии сложных или нестандартных искажений.

2.Обоснована актуальность применения для решения задачи восстновления методов глубокого обучения, в том числе глубоких нейронных сетей сверточного и трансформерного типов, обладающих высокой обобщающей способностью и универсальностью. Показано, что основным фактором, ограничивающим эффективность нейросетевых алгоритмов, является нехватка репрезентативных обучающих данных. В связи с этим выделена важная роль средств аугментации данных как ключевого инструмента расширения обучающих наборов и моделирования различных типов шумов и искажений.

3.Проанализированы существующие эвристические и генеративные методы аугментации, включая подходы на основе GAN, VAE и диффузионных моделей. Установлено, что современные генеративные методы позволяют достоверно синтезировать искажения, близкие к реальным, включая атмосферные осадки, шумы и артефакты. Отмечена важность использования специализированных архитектур, обеспечивающих управляемый перенос стилей и устойчивость к разнообразию входных данных.

4.Сформулирована необходимость комплексного подхода, объединяющего нейросетевые алгоритмы восстановления изображений с целенаправленной

аугментацией данных, как основа для построения более эффективных алгоритмов в условиях ограниченности данных и высокой вариативности шумовых воздействий. Предложена общая схема исследований, реализующая данный подход.

2. Теоретические обоснования возможных способов модификации механизма внимания для обеспечения регуляризации процесса обучения в нейронных сетях-трансформерах

В виду наличия описанных выше определенных трудностей при реализации механизма внимания в моделях-трансформерах во многих работах исследуются различные модификации стандартного варианта. Авторы XCiT (Cross-Covariance Image Transformers) [72] используют транспонированное внимание. Вычислительная сложность, которого квадратично зависит не от числа патчей, а от размерности эмбеддинга (канальной информации). Для учета увеличения взаимодействия между соседними патчами авторы добавляют два дополнительных сверточных слоя 3x3. Данный метод применим к задачам, где важны локальные взаимодействия, но не подходит для обработки изображений с сильной глобальной зависимостью между пикселями.

Авторы PS-ViT (Pooling and Attention Sharing) [73]. для патчей используют пуллинг с целью уменьшения размерности, с каждым слоем размерность изображения уменьшается в два раза. Помимо этого, применяется Attention Sharing – одинаковая матрица внимания для нескольких подряд идущих слоев. Данный метод лучше всего работает в задачах, где высокая детализация некритична, из-за этого возникают трудности при обработки текстурных изображений.

VOLO (Vision Outlooker for Visual Recognition) [74]. В методе отсутствует стандартный механизм внимания, но есть линейный слой, который преобразует входные данные в матрицу внимания. Values (V) в механизме внимания считаются как обычно. Матрица внимания состоит из сверток, получается гибрид сверточного слоя и стандартного self-attention. Данный метод эффективен в задачах классификации и распознавания, но может быть недостаточно точен для улучшения качества изображений.

SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution [75]. Модель SwinFIR – это недавняя вариация

модели SwinIR, которая использует компоненты быстрой свертки Фурье (FFC) для извлечения глобальной информации, подходящей для задачи сверхразрешения. Данный метод ресурсозатратен. Кроме того, авторами не приводятся результаты в области улучшения качества изображений.

Comprehensive and Delicate: An Efficient Transformer for Image Restoration [76]. Авторы используют суперпиксели для механизма внимания. Помимо обычных сверток, применяются и групповые. Кроме выделения суперпикселей в алгоритме используется канальное и пространственное внимание. Результаты от них в дальнейшем объединяются в единую карту признаков по разработанному авторами алгоритму объединения (Dual Adaptive Neural Block). Данный алгоритм сложен в настройке гиперпараметров, поэтому его трудно адаптировать под различные задачи компьютерного зрения.

DAT++: Spatially Dynamic Vision Transformer with Deformable Attention [77]. В статье авторы используют матрицу смещений в дополнение к механизму внимания. Матрица смещений зависит от query (Q) и прибавляется к текущим позициям патчей. Следовательно, для расчёта карт внимания используется не окно KxK, а другая область, зависящая от матрицы смещений.

Key-Graph Transformer for Image Restoration [78]. В данной работе применяется глобальный механизм внимания, но сравнивается только фиксированное число пикселей, которые имеют наибольшую корреляцию по входным значениям.

How Powerful Potential of Attention on Image Restoration [79]. Авторы предлагают убрать блоки линейного погружения. Вместо этого они используют три взаимосвязанных блока внимания с уменьшением пространственных размерностей входных данных в два раза для каждого следующего слоя.

Activating More Pixels in Image Super-Resolution Transformer [80]. Авторы добавляют канальный механизм внимания. Однако берут его из сверточных нейронных сетей (без K, Q, V) и прибавляют с некоторым коэффициентом к результату классического пространственного внимания трансформеров. Помимо этого, авторы группируют слои с механизмом внимания и в конце каждой группы

добавляют отдельный слой внимания с перекрытием окон (похожий на Swin). Однако это увеличивает сложность алгоритма, так как помимо пространственной информации надо учитывать и канальную.

Prompt-based Ingredient-Oriented All-in-One Image Restoration [81]. Авторы модифицируют механизм внимания обычной аддитивной составляющей к K, Q, V до функции softmax. Эта составляющая – обучаемый параметр. Но считается, что это уже Prompt Learning. С их специфичной архитектурой нужно дообучать только данную аддитивную составляющую, а не все слои, чтобы нейронная сеть научилась восстанавливать изображения от совершенно другого типа искажений. Для данного метода необходимо заранее знать априорную информацию об искажениях изображения.

2.1. Схема вычисления механизма внимания (самовнимания) применительно к задаче восстановления изображений

Общая схема вычисления механизма внимания в нейронных сетях – трансформерах представлена на рисунке 2.1. Представленная схема наглядно иллюстрирует механизм внимания в моделях-трансформерах применительно к задаче восстановления изображений.

На первом этапе входной тензор, представляющий изображение в виде последовательности патчей или пикселей, пропускается через три независимых линейных слоя, формируя три набора признаков: запросы (Q), ключи (K) и значения (V). Каждый из этих тензоров далее преобразуется в форму, подходящую для многоголового внимания, где каждая голова обрабатывает только свою часть признакового пространства.

Затем происходит вычисление матрицы внимания: каждый запрос сравнивается с каждым ключом, формируя карту весов внимания. Эта карта нормализуется с помощью активации softmax, после чего используется для взвешенного суммирования соответствующих значений. Для дальнейшего математического анализа необходимо детально расписать механизм внимания в моделях-трансформерах.

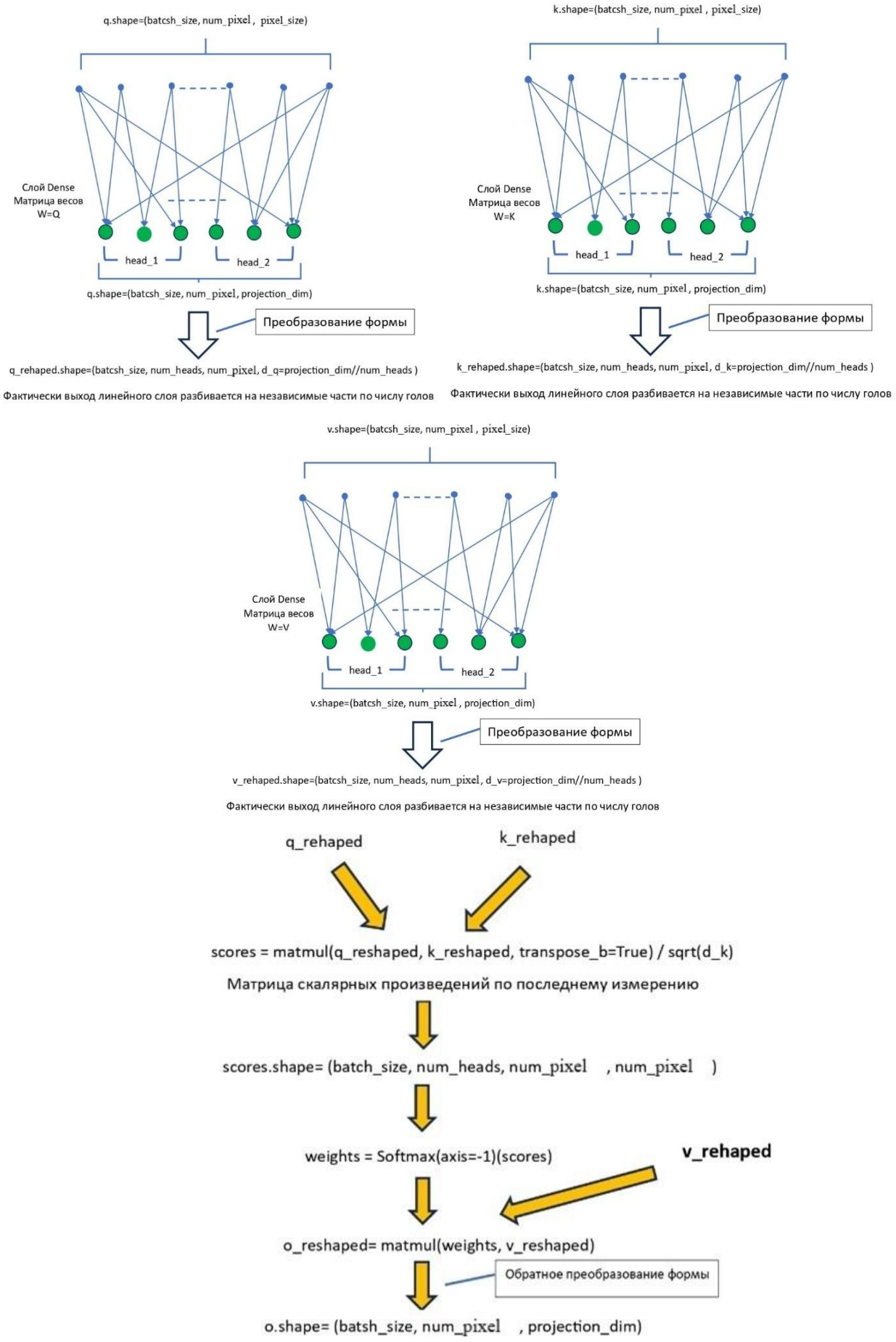


Рисунок 2.1 – Схема тензорного потока для механизма внимания

Пусть в соответствии с принятой терминологией [82] $q=(q_1,...,q_p)^T$ – вектор запроса и $k=(k_1,...,k_s)^T$ – вектор ключа. Сравнить их можно, отобразив в общее пространство погружения размерности d, для чего вводят обучаемые матрицы W_q, W_k, где $\mathrm{W}_q \in \mathbf{R}^{d\times p}, W_k \in \mathbf{R}^{d\times s}$. В варианте самовнимания (self-attention, SA): $q=k,\ p=s=m$. Пусть также $v=(v_1,...,v_v)^T$ – вектор значений. В общем случае он также подвергается своему линейному погружению. Запрос и ключи после выполнения погружения будем обозначать как $\hat{q},\hat{k}\in\mathbf{R}^d$. Также для простоты выкладок условимся, что весь последующий анализ будет проводится для одной «головы» с потенциальным переносом результатов на случай «многоголового внимания».

Будем считать, что это запросы и ключи случайные величины с независимыми компонентами с нулевым средним и единичной дисперсией, то среднее их скалярного произведения равно 0, а дисперсия равна d. Чтобы дисперсия скалярного произведения оставалась равной 1 независимо от размера входов, принято делить на $\sqrt{d}$. Тогда SA можно вычислить в форме масштабированного скалярного произведения:

$$a(\hat{q},\hat{k}) = \hat{q}^T\hat{k} / \sqrt{d},$$

где $\sqrt{d}$ – масштабный коэффициент, обеспечивающий фиксированную дисперсию скалярного произведения (равной единице для гауссовских центрированных величин) независимо от размера $\hat{q},\hat{k}\in\mathbf{R}^d$.

На практике при восстановлении изображений используется скалярное произведение между пикселями одного изображения и $Q\in\mathbf{R}^{n\times d}, K\in\mathbf{R}^{n\times d}, \mathrm{V}\in\mathbf{R}^{n\times v}$, где n – число пикселей изображения, используемое для расчета матрицы внимания, а d – размер признакового пространства. В целях упрощения дальнейших обозначений для элементов строк и столбцов матриц, связанных с $\hat{q},\hat{k}$, верхний надстрочный символ использовать не будем. Тогда, согласно [82], можно вычислить взвешенные вниманием выходы следующим образом:

$$attn(Q,K,V)=\text{softmax}\left(QK^T/\sqrt{d}\right)V\in\mathbf{R}^{n\times v},$$

$$\mathrm{Q}=\begin{pmatrix} q_1^{(1)} & \cdots & q_d^{(1)} \\ \vdots & \vdots & \vdots \\ q_1^{(n)} & \cdots & q_d^{(n)} \end{pmatrix}=\begin{pmatrix} q^{(1)} \\ \vdots \\ q^{(n)} \end{pmatrix},\ K=\begin{pmatrix} k_1^{(1)} & \cdots & k_d^{(1)} \\ \vdots & \vdots & \vdots \\ k_1^{(n)} & \cdots & k_d^{(n)} \end{pmatrix}=\begin{pmatrix} k^{(1)} \\ \vdots \\ k^{(n)} \end{pmatrix},$$

$$V=\begin{pmatrix} v_1^{(1)} & \cdots & v_v^{(1)} \\ \vdots & \vdots & \vdots \\ v_1^{(n)} & \cdots & v_v^{(n)} \end{pmatrix}=\begin{pmatrix} v^{(1)} \\ \vdots \\ v^{(n)} \end{pmatrix},\ (2.1)$$

где $\text{soft}\max(...)$ – функция softmax, применяемая для матрицы построчно.

Здесь и далее верхний индекс определяет номер строки, нижний индекс – номер столбца и обозначения элементов строк и столбцов проводится следующим образом: $q^{(j)}=(q_1^{(j)},...,q_d^{(j)}),k^{(j)}=(k_1^{(j)},...,k_d^{(j)}),\ k_i=(k_1^{(i)},...,k_d^{(i)})^T,v^{(j)}=(v_1^{(j)},...,v_v^{(j)}),$ $v_i=(v_i^{(1)},...,v_i^{(n)})^T$.

Приведенные выше соотношения представим в виде

$$\mathrm{Q}K^T=\frac{1}{\sqrt{d}}\begin{pmatrix} q^{(1)}k_1 & q^{(1)}k_2 & \cdots & q^{(1)}k_n \\ q^{(2)}k_1 & q^{(2)}k_2 & \cdots & q^{(2)}k_n \\ \vdots & \vdots & \vdots & \vdots \\ q^{(n)}k_1 & q^{(n)}k_2 & \cdots & q^{(n)}k_n \end{pmatrix},\ q^{(i)}k_j=\frac{1}{\sqrt{d}}\sum_{p=1}^{d}q_p^{(i)}k_p^{(j)},\ \ i=\overline{1,n},\ j=\overline{1,n}.$$

Соответствующая матрица внимания после активации softmax построчно имеет вид:

$$A=\begin{pmatrix} \alpha_1^{(1)} & \cdots & \alpha_n^{(1)} \\ \vdots & \vdots & \vdots \\ \alpha_1^{(n)} & \cdots & \alpha_n^{(n)} \end{pmatrix},\ \ \alpha^{(i)}=(\alpha_1^{(i)},...,\alpha_n^{(i)})=\text{soft}\max\left\{\frac{q^{(i)}k_1}{\sqrt{d}},....,\frac{q^{(i)}k_n}{\sqrt{d}}\right\},i=\overline{1,n}.\ (2.2)$$

Т.е. фактически механизм внимания реализуется путем вычисления скалярных произведений запроса $q^{(i)}$ со всеми ключами, а механизм самовнимания – всех запросов со всеми другими запросами (пикселями). В итоге взвешенные значения вектора значений можно получить следующим образом:

$$O=\begin{pmatrix} o_1^{(1)} & \cdots & o_v^{(1)} \\ \vdots & \vdots & \vdots \\ o_1^{(n)} & \cdots & o_v^{(n)} \end{pmatrix}=\begin{pmatrix} o^{(1)} \\ \vdots \\ o^{(n)} \end{pmatrix}=AV=\begin{pmatrix} \alpha_1^{(1)} & \cdots & \alpha_n^{(1)} \\ \vdots & \vdots & \vdots \\ \alpha_1^{(n)} & \cdots & \alpha_n^{(n)} \end{pmatrix}\begin{pmatrix} v_1^{(1)} & \cdots & v_v^{(1)} \\ \vdots & \vdots & \vdots \\ v_1^{(n)} & \cdots & v_v^{(n)} \end{pmatrix}, \quad (2.3)$$

$$o^{(i)}=(o_1^{(i)},...,o_v^{(i)})=\alpha^{(i)}V\,,\;\; o_j^{(i)}=\alpha^{(i)}v_j=\sum_{p=1}^{n}\alpha_p^{(i)}v_j^{(p)},\;\; i=\overline{1,n},\; j=\overline{1,v}\,.$$

В (2.3) O – матрица выходов слоя внимания трансформера, каждая строка которой соответствует своему пикселю, в которой осуществляется взвешивание компонентов с одинаковыми номерами (j) для всех пикселей с номерами $i=\overline{1,n}$. Очевидно, что чем больше каждое скалярное произведение, тем больший вес после softmax получают соответствующие компоненты значений, которые взвешивается при использовании (2.1) - (2.3).

Анализ представленных соотношений показывает, что каждая строка произведения QK^T представляет собой набор скалярных произведений запросов и ключей различных пикселей. Например, первая строка $q^{(1)}K^T=(q^{(1)}k^{(1),T},q^{(1)}k^{(2),T},...,q^{(1)}k^{(n),T})^T$ представляет сравнение запроса и ключей (при самовнимании фактически все скалярные произведений векторов первого пикселя со всеми остальными):

$$q^{(1)}K^T=(q^{(1)}q^{(1),T},q^{(1)}q^{(2),T},...,q^{(1)}q^{(n),T})^T.$$

Очевидно, что чем больше каждое скалярное произведение (степень схожести патчей), тем больший вес на выходе softmax и тем больше вес получают соответствующие компоненты значений, которое взвешивается в формуле (3).

Таким образом, каждый элемент выходного тензора содержит агрегированную информацию из всех других позиций изображения, учитывая их взаимную значимость. На последнем этапе результаты от всех голов объединяются и приводятся к исходной размерности проекции. Для лучшего понимания механизма внимания, необходимо сначала понять особенности данного механизма и визуализировать его представление.

2.2. Исследование особенностей механизма внимания и его визуализация в задачах восстановления изображений

Исследователями [83-84] предпринимаются попытки визуализировать механизм внимания трансформера ViT. Однако для задач восстановления изображений с большим разрешением исследователи сталкиваются с проблемами вычислительной сложности при визуализации механизма глобального пространственного внимания. Локальный механизм внимания позволяет решить проблему вычислительной сложности, но при этом теряется точность восстановления изображений и наглядность его интерпретации.

Помимо этого, ряд исследователей полагают, что знание априорной информации об изображении позволяет существенно улучшить результат восстановления [85]. В работе автора [86] подробно исследуется влияния априорной информации о силе аддитивного гауссовского шума на результат восстановления. Показывается, что добавление априорной информации об искажении способствует увеличению сходимости нейронной сети и улучшает результат восстановления изображений. Однако вопрос о взаимосвязи механизма внимания с априорной информацией об изображении остается открытым до сих пор.

Понимание принципов действия механизма внимания в моделях трансформерах при улучшении качества изображений позволит оптимизировать карты внимания и уменьшить избыточность модели. Также это дает возможность производить постобработку карт внимания, сделать их полностью настраиваемыми, как предлагают исследователи для сверточных нейронных сетей в работе [87].

В настоящей работе берется за основу стандартная модель-трансформер [36] для наглядности визуализации канальных карт внимания. Используется глобальный механизм внимания, основанный на канальных картах, предлагая способ определения их степени важности при улучшении качества изображения.

В работе для исследования механизма внимания решено было использовать следующие типы помех и искажений: аддитивный гауссовский шум, размытие,

засветку изображения с одной из сторон, мелкие аппликативные помехи, шум соль и перец, алгоритмы генерации, который показаны в работе [40]. Для проверки результатов экспериментов использовалась проверочная выборка из датасета ImageNet, на него же и накладывались перечисленные выше типы помех. Также решено было использовать датасет SIDD с реальными типами искажений, чтобы подтвердить гипотезы, поставленные на искусственно сгенерированных помехах.

Для сравнения качества изображений решено было использовать метрики пиковое отношение сигнала к шуму (PSNR) и структурного подобия (SSIM). Помимо этого, применялся визуальный анализ изображений, так как исследователями [88] было показано, что данные метрики не всегда адекватно соотносятся с качеством изображений.

Для выявления закономерностей между априорной информацией об изображении и его канальными картами внимания были выделены следующие его параметры: контрастность, измеряемая как среднеквадратическое отклонение между пикселями изображения, контрастность Майкельсона, средняя амплитуда значений и дисперсия изображения после применения к нему высокочастотного фильтра Лапласа. Контрастность Майкельсона определялась как:

$$\gamma = \frac{I_{\max} - \mathrm{I}_{\min}}{I_{\max} + \mathrm{I}_{\min}},$$

где *Imin, Imax* – минимум и максимум интенсивности пикселей изображения.

Затем проводились исследования о влиянии механизма внимания на результат восстановления изображений. Было проведено шесть экспериментов, в ходе которых сделаны определенные выводы, подробно представленные в работе автора [89]. В качестве исходных данных использовались изображения из датасетов ImageNet и SIDD.

1. **Сравнение распределений карт внимания, полученных от эталонных и зашумленных изображений.** Для сравнения распределений решено было использовать t-тест на равенство средних, ввиду особой значимости амплитуды карт внимания. Было выявлено, что по каждому из параметров нулевая

гипотеза о равенстве средних не подтвердилась, что показало статистического различие распределений карт внимания зашумленных и эталонных изображений.

2. **Проверка наличия значимого изменения карт внимания при наложении мелких аппликативных помех.** Данное утверждение решено было подтвердить путем обучения простой полносвязной нейронной сети. На ее вход подавались карты внимания зашумленных и эталонных изображений. Нейронная сеть училась определять какие из карт внимания получены от эталонных изображений, а какие нет. Обученный классификатор имел точность свыше 99 % на заранее отобранной тестовой выборке. Это показало различие между картами внимания эталонных и искаженных изображений.

3. **Исследование наличия зависимости RMSE по картам внимания от вида помех.** В этом эксперименте выбиралось 500 эталонных изображений из датасета ImageNet, затем производилось их зашумление каждым из видов помех. После чего считался RMSE для карт внимания каждого изображения. Затем производилось его усреднение по всем изображениям. Расчет RMSE для каждого изображения производился по следующей формуле:

$$aRMSE = \sqrt{\frac{1}{C^2}\sum_i \sum_j (Ax_{ij} - Ay_{ij})^2},$$

где C – размерность канальной карты внимания; A_x, A_y – матрицы внимания размерами $C \times C$, полученные от изображения, зашумлённого помехой типа *x* и типа *y* соответственно.

Средний RMSE считался как среднее по всем *aRMSE,* полученных по канальным картам внимания:

$$RMSE = \frac{\sum_i^N aRMSE_i}{N}.$$

В таблице 2.1 представлены результаты *RMSE* для первого слоя с механизмом внимания. Таблица симметричная, поэтому показана только её часть выше главной диагонали.

Таблица 2.1 – RMSE по типам помех для первого слоя с механизмом внимания

Тип помех	Clear	Gauss	S&P	Applicative	Засветка	Размытие
Clear	0.0	0.184	0.183	0.196	0.20	**0.26**
Gauss	-	0.0	0.013	0.059	0.147	**0.193**
S&P	-	-	0.0	0.06	0.15	**0.195**
Applicative	-	-	-	0.0	0.124	**0.155**
Засветка	-	-	-	-	0.0	**0.157**
Размытие	-	-	-	-	-	0.0

Из таблицы 2.1 можно сделать вывод, что карты внимания, полученные от изображений зашумленных аддитивным гауссовским шумом и шумом соль и перец, имеют наименьшее различие, ближе к ним идут мелкие аппликативные помехи. Засветка изображения и размытие – существенно отличаются от всех остальных. Также стоит отметить, что карты внимания, полученные на размытых изображений, наиболее существенно отличаются от эталонных. По данному выводу можно судить, что механизм внимания чувствителен не только к пространственной структуре помехи, но и к её типу. В таблице 2.2 представлены аналогичные значения, но только для последнего слоя с механизмом внимания.

Таблица 2.2 – RMSE по типам помех для последнего слоя с механизмом внимания

Тип помех	Clear	Gauss	S&P	Applicative	Засветка	Размытие
Clear	0.0	0.206	0.197	0.19	0.19	**0.278**
Gauss	-	0.0	0.077	0.168	0.184	**0.307**
S&P	-	-	0.0	0.162	0.164	**0.298**
Applicative	-	-	-	0.0	0.149	**0.236**
Засветка	-	-	-	-	0.0	**0.258**
Размытие	-	-	-	-	-	0.0

Стоит отметить, что RMSE в таблице 2.2 отличается от результатов, показанных выше в таблице 2.1, но закономерность не изменилась по каждому из

типов помех. Однако во всех случаях средний RMSE по изображениям значительно вырос. Анализируя карты внимания на промежуточных слоях, нами был сделан вывод, что различие между картами внимания, полученных от изображений с различными типами помех, только возрастает к выходу нейронной сети. Что в целом доказывает, что нейронная сеть на первых слоях извлекает простые пространственные и яркостные признаки из изображения, а в конце уже смотрит на семантически более сложные признаки, позволяющие различать одну помеху от другой.

4. **Анализ влияния параметров изображения (яркость и контрастность) на карты внимания.** При проведении такого анализа решено было разбить изображения на 3 класса. Для этого использовался алгоритм кластеризации *K-Means*. По результатам кластеризации было выявлено три явных класса. В первый класс попали изображения с высокой степенью контрастности, во второй класс попали изображения с низкой степенью контрастности, а в третий – со средней степенью контрастности, но с большим значением средней амплитуды фильтра Лапласа, что говорит о наличии больших текстурных областей на изображениях. Визуальный анализ изображений подтвердил эти выводы. Затем карты внимания эталонных изображений, полученные от обученной модели трансформер, были распределены по соответствующим кластерам. Статистический анализ показал, что нет никакой зависимости и существенных различий между кластерами на уровне карт внимания. Отсюда можно сделать вывод, что на механизм внимания указанные параметры изображений существенно не влияют.

5. **Выявление локальной взаимосвязь между частями изображения и его картой внимания.** В этом эксперименте был использован подход, описанный в работе [83]. При пропускании изображения через нейронную сеть часть карты внимания маскировалась (занулялась), а затем визуализировался результат на выходе. Было выбрана маска размером в четверть канальной карты внимания. Затем она сдвигалось с постоянным шагом по карте внимания. Шаг был равен $C/6$, где CC – размер карты внимания. После чего измерялись

метрики PSNR и SSIM для определения качества восстановленных изображений. Результаты подобных действия показаны на рисунке 2.2.

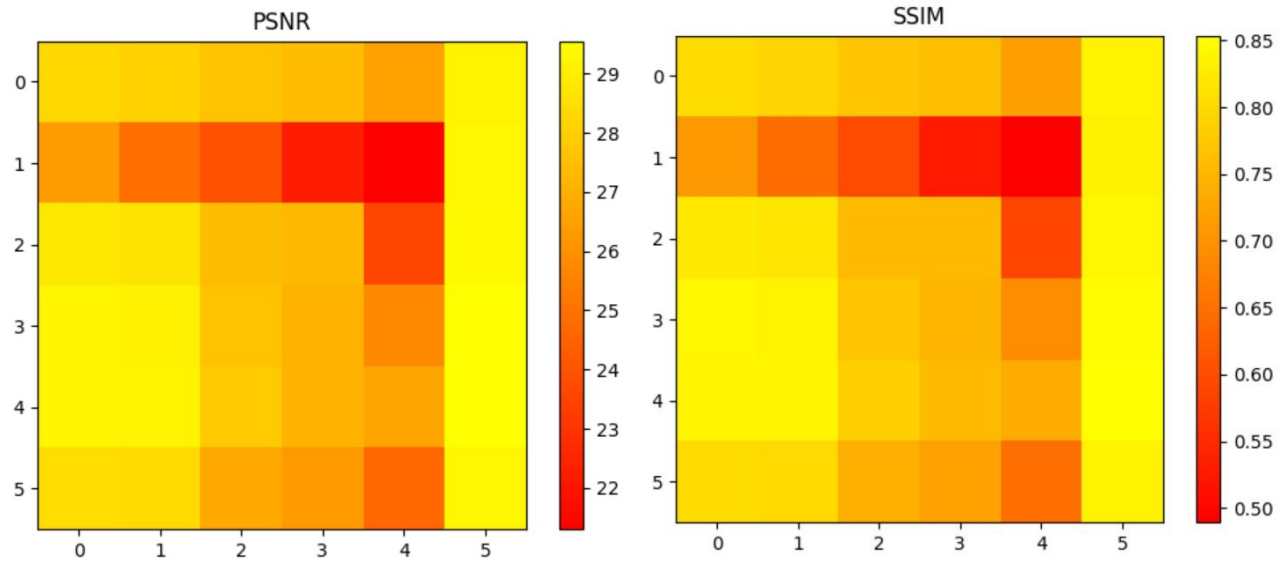


Рисунок 2.2 – Изображение тепловой карты PNSR и SSIM после зануления части карты внимания

Стоит отметить, что тепловые карты, полученные для PSNR и SSIM – схожи. Изображения наихудшего качества получаются при зануленииии центральной части карт внимания. Отсюда следует, что она вносит больший вклад в восстановление изображений, что подчеркивает неоднородность карт внимания.

Также было выявлено, что последние слои с механизмом внимания оказывают меньшее влияние на восстанавливаемое изображение, чем предыдущие. Если проводить зануление карты внимания на первых слоях, то результат восстановления значительно ухудшится.

6. **Исследование зависимости качества изображения от размера маскируемой области карты внимания.** Этом эксперимент проводился по схожей схеме, что и предыдущий. Примеры результатов восстановления, полученные в этих условиях, представлены на рисунке 2.3.

a)

b)

c)

d)

Рисунок 2.3 – Восстановленные изображения при маскировании: a) – 0 % карты внимания, b) – 25%, c) – 50%, d) – 100 % (исходное зашумленное изображение)

При маскировании более 50% карты внимания качество восстановления сильно падает. Результат практически аналогичен исходному зашумлённому изображению. При маскировании 25 % карты внимания качество изображения также падает, но не настолько сильно: видны всего лишь мелкие артефакты от исходного шума. Эти выводы подтвердились на любом изображении независимо от типа помех.

Все представленные выше эксперименты проведены для канального механизма внимания, тоже самое можно повторить и для пространственного,

позволив понять его особенности и важность в процессе восстановления изображений.

Для большей интерпретации пространственного механизма внимания визуализируем его карты. В качестве архитектуры для исследования пространственного механизма внимания была выбрана SwinIR [30] используется локальное внимание и механизм сдвига окон – черные полосы на изображения – эффект от него.

Ниже приводятся матрицы коэффициентов механизма внимания на разных стадиях обучения. Так как в исходной работе используется архитектура с локальным механизмом внимания для окна $N \times N$, то предлагается следующий алгоритм формирования визуализаций для карт внимания:

1) Объединить все патчи в одно изображение.
2) Возьмем матрицу для первой головы.
3) Нормируем все значения от 0 до 1.
4) Отобразим полученную матрицу, шахматный артефакт на рисунках как раз показывает места склейки карт внимания для патчей изображения.

На рисунке 2.4 показано соответствие между исходным изображение и результатом визуализацией карты внимания.

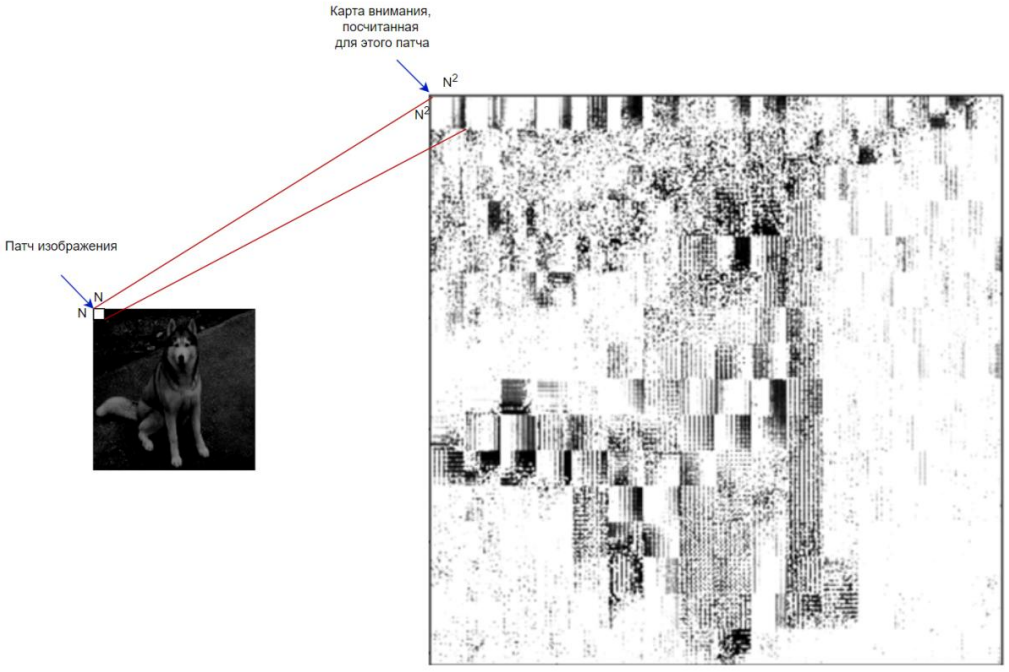
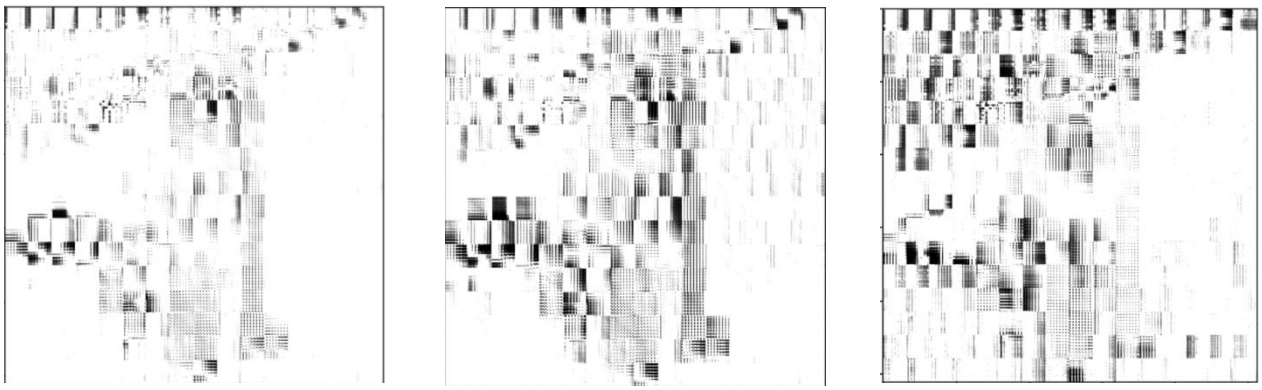


Рисунок 2.4 – Соотношение исходного изображения с матрицей внимания

На рисунке 2.5 показаны карты внимания без модификаций в зависимости от числа итераций при обучении.



a) *b)* *c)*

Рисунок 2.5 – Визуализация механизма внимания: a) – карты внимания после 50 тысяч итераций при обучении, b) – после 10 тысяч итераций при обучении, c) после 5 тысяч итераций при обучении.

На основе представленных выше рисунков можно сделать вывод, что, чем больше обучается нейронная сеть трансформер, тем больше карты его внимания похожи на контуры исходного изображения. Визуальный анализ подтверждает факт структурной предвзятости механизма внимания [3], позволяющий эффективно восстанавливать изображения на обучающей выборке, но теряющий качество на реальных.

2.3. Теоретические обоснования возможных способов структурной регуляризации механизма самовнимания на основе внесения стохастических составляющих различных типов

Одним из важных направлений является исследование возможностей совершенствования механизма самовнимания моделей трансформеров на основе использования методов структурной регуляризации, что потенциально позволяет снизить эффекты, связанные с переобучением. Структурная регуляризации, или регуляризация путем ограничений на структуру модели предполагает внесение в механизм SA стохастических воздействий, направленных на либо на случайное отключение нейронных связей (классический dropout), исключение части

размерностей эмбеддингов (DropDim), исключение части голов внимания (DropHead), стохастическое исключение весов внимания (DropAttention, DropKey) и пр. [90-93].

Данная работа является развитием и обобщением исследований [92-93] и направлена на обоснование и исследование метода регуляризации путем внесения мультипликативной стохастической составляющей с произвольным дискретным или непрерывным распределением для весовых коэффициентов SA.

В соответствии с соотношениями (2.1-2.3) матрица внимания после активации softmax построчно имеет вид:

$$A=\begin{pmatrix} \alpha_1^{(1)} & \cdots & \alpha_n^{(1)} \\ \vdots & \vdots & \vdots \\ \alpha_1^{(n)} & \cdots & \alpha_n^{(n)} \end{pmatrix},\quad \alpha^{(i)}=(\alpha_1^{(i)},...,\alpha_n^{(i)})=\text{soft}\max\left\{\frac{q^{(i)}k_1}{\sqrt{d}},....,\frac{q^{(i)}k_n}{\sqrt{d}}\right\},i=\overline{1,n}$$

2.3.1. Структурная регуляризация путем внесения мультипликативной составляющей

Основной идеей структурной регуляризации весов внимания состоит в выполнении сглаживания строк матрицы весов внимания, таким образом, чтобы исключить ситуацию, когда выходные данные модели будут контролироваться несколькими разреженными блокам и, наоборот, обеспечить более плавное распределение α [92-93]. Рассмотрим строгое доказательство возникновения эффекта сглаживания при внесении мультипликативной стохастической составляющей для элементов матрицы выхода модуля внимания в общем случае.

Основное доказательство. Пусть имеется случайная величина $\mathbf{h}^{(i)}$, вносящая независимое случайное воздействие на элементы строки выхода $o^{(i)}$ взвешенной softmax матрицы внимания. Ее значения в позиции p будем обозначать как $h_p^{(i)}$. Тогда внесение мультипликативной составляющей в интересах регуляризации матрицы внимания может осуществляться на основе следующих соотношений:

$$o_j^{(i)} = \sum_{p=1}^{n}\left(\frac{h_p^{(i)}\alpha_p^{(i)}v_j^{(p)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\right),\ j=\overline{1,\nu}\ , \tag{2.4}$$

$$o^{(i)} = \sum_{p=1}^{n}\left(\frac{h^{(i)}\otimes\alpha^{(i)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\right)\left(v_1,...v_v\right) = \sum_{j=1}^{v}p^{(i)}v_j\,,\ h^{(i)} = (h_1^{(i)},...,h_n^{(i)})\,,$$

$$p^{(i)} = \frac{(h_1^{(i)}a_p^{(i)},...,h_n^{(i)}a_p^{(i)})}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}.$$

Пусть для определенности $\mathbf{h}^{(i)}$ принимает дискретный ряд значений на интервале [0,1] $\mathbf{h}^{(i)}\in\{I_0,...I_m\},\ I_e = e/m,\ e=\overline{0,\mathrm{m}}$ с распределением $P_p\left(I_e\right) = P_h(h_p^{(i)} = I_e),\ p=\overline{1,n},\ e=\overline{0,m}$. Пусть также исходные веса внимания в двух позициях имеют соотношение $\alpha_t^{(i)} > a_s^{(i)}$. Оценим в этих условиях соотношение условных математических ожиданий величин $h_t^{(i)}, h_s^{(i)}$, осуществляющих коррекцию исходных весов внимания на основе следующих соотношений:

$$\Delta c_{t,s}^{(i)} = c_t^{(i)} - c_s^{(i)}\,, \tag{2.5}$$

$$c_t^{(i)} = M\left[\frac{h_t^{(i)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\middle/\left\{\alpha_p^{(i)}\right\},\left\{h_k^{(i)},k\neq t,s\right\}\right] = M_{a,h}\left[\frac{h_t^{(i)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\right] =$$

$$= \sum_{e=0}^{m}\sum_{q=0}^{m}\frac{I_e}{\sum_{k=1,k\neq t,s}^{n}h_k^{(i)}\alpha_k^{(i)} + I_e\alpha_t^{(i)} + I_q\alpha_s^{(i)}}P_t(I_e)P_s(I_q)\,,$$

$$c_s^{(i)} = M\left[\frac{h_s^{(i)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\middle/\left\{\alpha_p^{(i)}\right\},\left\{h_k^{(i)},k\neq t,s\right\}\right] = M_{a,h}\left[\frac{h_s^{(i)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\right] =$$

$$= \sum_{e=0}^{m}\sum_{q=0}^{m}\frac{I_q}{\sum_{k=1,k\neq t,s}^{n}h_k^{(i)}\alpha_k^{(i)} + I_e\alpha_t^{(i)} + I_q\alpha_s^{(i)}}P_t(I_e)P_s(I_q)\,.$$

Здесь I_e, I_q значения $h_t^{(i)}\ h_s^{(i)}$ соответственно. Введем матрицу попарных сочетаний индексов

$$\Omega = \begin{pmatrix} (I_0, I_0) & (I_0, I_1) & \dots & (I_0, I_m) \\ (I_1, I_0) & (I_1, I_1) & & (I_1, I_m) \\ (I_2, I_0) & (I_2, I_1) & \dots & (I_2, I_m) \\ \vdots & \vdots & \vdots & \vdots \\ (I_m, I_0) & (I_m, I_1) & & (I_m, I_m) \end{pmatrix}.$$

Выполним перегруппировку слагаемых в (2.5) таким образом, чтобы в $\Delta c_{t,s}^{(i)} = c_t^{(i)} - c_s^{(i)}$ по соседству находились составляющие с симметричными относительно главной диагонали матрицы Ω позициями. Для сокращения объема формульных описаний введем обозначения:

$$E_{e,e-r} = \left[\frac{I_e}{\sum_{k=1,k\neq t,s}^{n} h_k^{(i)} \alpha_k^{(i)} + I_e \alpha_t^{(i)} + I_{e-r} \alpha_s^{(i)}} - \frac{I_{e-r}}{\sum_{k=1,k\neq t,s}^{n} h_k^{(i)} \alpha_k^{(i)} + I_e \alpha_t^{(i)} + I_{e-r} \alpha_s^{(i)}} \right] P_t(I_e) P_s(I_{e-r}),$$

$$E_{e-r,e} = \left[\frac{I_{e-r}}{\sum_{k=1,k\neq t,s}^{n} h_k^{(i)} \alpha_k^{(i)} + I_{e-r} \alpha_t^{(i)} + I_e \alpha_s^{(i)}} - \frac{I_e}{\sum_{k=1,k\neq t,s}^{n} h_k^{(i)} \alpha_k^{(i)} + I_{e-r} \alpha_t^{(i)} + I_e \alpha_s^{(i)}} \right] P_t(I_{e-r}) P_s(I_e).$$

Тогда исходное выражение (2.5) можно представить в виде:

$$\Delta c_{t,s}^{(i)} = \sum_{e=0}^{m} E_{e,e} + \sum_{e=1}^{m} \left\{ E_{e,e-1} + E_{e-1,e} \right\} + \sum_{e=2}^{m} \left\{ E_{e,e-2} + E_{e-2,e} \right\} + \dots \qquad (2.6)$$

$$\dots \sum_{e=m-1}^{m} \left\{ E_{e,e-m+2} + E_{e-m+2,e} \right\} + \left\{ E_{m,0} + E_{0,m} \right\} = \sum_{u=0}^{m} \sum_{e=u}^{m} \left\{ E_{e,e-u} + E_{e-u,e} \right\}.$$

Рассмотрим каждую составляющую в сумме с симметричными позициями в матрице Ω

$$E_{e,e-r} = \frac{\delta r}{\sum_{k=1,k\neq t,s}^{n} h_k^{(i)} \alpha_k^{(i)} + I_e \alpha_t^{(i)} + I_{e-r} \alpha_s^{(i)}} P_t(I_e) P_s(I_{e-r}), \qquad (2.7)$$

$$E_{e-r,e} = \frac{-\delta r}{\sum_{k=1,k\neq t,s}^{n} h_k^{(i)} \alpha_k^{(i)} + I_{e-r} \alpha_t^{(i)} + I_e \alpha_s^{(i)}} P_t(I_{e-r}) P_s(I_e) \ , \ \delta r = r/m, r = \overline{0,m}$$

Очевидно, что для любого произвольного произведения $P_t(I_{e-r}) P_s(I_e) = P_t(I_e) P_s(I_{e-r}) = P(I_e) P(I_{e-r})$, если $P_p(I_e) = P(I_e),\ p = \overline{1,n}$, т.е. распределение случайной составляющей не зависит от позиции в строке матрицы

внимания. Для дальнейшего анализа соотношения $E_{e,e-r}$ и $E_{e-r,e}$ в (7), учтем базовое неравенство $\alpha_t^{(i)} > a_s^{(i)}$, тогда для любого $r = \overline{0,m}$ получим:

$$\delta r \alpha_t^{(i)} = (I_e - I_{e-r}) a_t^{(i)} > \delta r a_s^{(i)} = (I_e - I_{e-r}) a_s^{(i)}, \quad I_e \alpha_t^{(i)} + I_{e-r} \alpha_s^{(i)} > I_{e-r} \alpha_t^{(i)} + I_e \alpha_s^{(i)}.$$

Таким образом, знаменатель для $E_{e,e-r}$ будет всегда больше $E_{e-r,e}$ в (2.7) и соответственно для $r = \overline{0,m}$ в (2.6)

$$E_{e,e-r} + E_{e-r,e} < 0. \tag{2.8}$$

Отметим, что для $\delta r = 0$ первая сумма в (2.6) $\sum_{e=0}^{m} E_{e,e} = 0$. Из (2.8) окончательно следует, что

$$\Delta c_{t,s}^{(i)} < 0, \quad c_t^{(i)} < c_s^{(i)}. \tag{2.9}$$

Напомним, что усреднение выполнялось для условных математических ожиданий

$$c_{t,s}^{(i)} = M\left[\frac{h_{t,s}^{(i)}}{\sum_{k=1}^{n} h_k^{(i)} \alpha_k^{(i)}} \middle/ \left\{ \alpha_p^{(i)} \right\}, \left\{ h_k^{(i)}, k \neq t,s \right\} \right] = M_{a,h}\left[\frac{h_{t,s}^{(i)}}{\sum_{k=1}^{n} h_k^{(i)} \alpha_k^{(i)}} \right].$$

Учитывая, что неравенство $c_t^{(i)} < c_s^{(i)}$ выполняется для любой произвольной комбинации остальных значений $\mathbf{h^{(i)}}$: $\left\{ h_k^{(i)}, k \neq t,s \right\}$, нетрудно видеть, что оно также будет верно для безусловных математических ожиданий, определяемых путем взвешенного суммирования по всем возможным комбинациям:

$$\overline{c}_t^{(i)} < \overline{c}_s^{(i)}, \quad \overline{c}_{t,s}^{(i)} = \sum_{\left\{ h_k^{(i)} \in I_\Sigma, k \neq t,s \right\}} c_{t,s}^{(i)} \prod_{\substack{r=1 \\ r \neq t,s}}^{n} P_r(I_e^{(r)}),$$

где $I_e^{(r)}$ – значение $h_r^{(i)}$ в позиции $r \neq t,s$ в конкретной комбинации из множества возможных $\left\{ h_k^{(i)}, k \neq t,s \right\}$. Представленные обоснования означают, что внесение мультипликативной составляющей обладает регуляризационным эффектом по отношению к росту весовых коэффициентам матрицы внимания. В среднестатистическом смысле это влияние проявляется в сглаживании весов внимания, снижая веса, имеющие чрезмерно большие значения. Тем самым, в

неявном виде, производится регуляризация процесса обучения, так как теперь выход каждого пикселя имеет вид:

$$\hat{o}_j^{(i)} = M_a[o_j^{(i)}] = M_h\left[\sum_{p=1}^{n}\left(\frac{h_p^{(i)}\alpha_p^{(i)}v_j^{(p)}}{\sum_{k=1}^{n}h_k^{(i)}\alpha_k^{(i)}}\right)\right] = \sum_{p=1}^{n}c_p^{(i)}\alpha_p^{(i)}v_j^{(p)}\,.$$

Следствие 1. Очевидно, также, что подобные неравенства могут быть получены для случая, когда значения $\mathbf{h}^{(i)} \in [0,1]$ принадлежат множеству континуум на заданном интервале (не обязательно в указанных границах) для которого задана непрерывная плотность распределения $P_h(h_p^{(i)})$, $p = \overline{1,n}$. Для этого достаточно представить:

$$\Delta c_{t,s}^{(i)} = c_t^{(i)} - c_s^{(i)} = \int_0^1\int_0^1\left(\frac{x}{\sum_{k=1,k\neq t,s}^{n}h_k^{(i)}\alpha_k^{(i)} + x\alpha_t^{(i)} + y\alpha_s^{(i)}} - \frac{y}{\sum_{k=1,k\neq t,s}^{n}h_k^{(i)}\alpha_k^{(i)} + x\alpha_t^{(i)} + y\alpha_s^{(i)}}\right)P_h(x)P_h(y)dxdy$$

и далее:

$$\Delta c_{t,s}^{(i)} = c_t^{(i)} - c_s^{(i)} = \int_0^1 P_h(x)dx\int_0^x\left(\frac{x}{R + x\alpha_t^{(i)} + y\alpha_s^{(i)}} - \frac{y}{R + x\alpha_t^{(i)} + y\alpha_s^{(i)}}\right)P_h(y)dy +$$

$$+\int_0^1 P_h(x)dx\int_x^1\left(\frac{x}{R + x\alpha_t^{(i)} + y\alpha_s^{(i)}} - \frac{y}{R + x\alpha_t^{(i)} + y\alpha_s^{(i)}}\right)P_h(y)dy,\quad R=\sum_{k=1,k\neq t,s}^{n}h_k^{(i)}\alpha_k^{(i)}\,.$$

Выполнив для первого и второго интегралов замену переменных следующим образом: $y = x - u,\ x = x,\ 0 \le \mathrm{u} \le x$, $x = y - u,\ y = y,\ 0 \le \mathrm{u} \le y$ (см. также рис.1) с учетом знаков якобианов преобразований, а также новых пределов интегрирования в первом интеграле для x: $a = 0,\ b = 1$ и для u: $a = 0,\ b = x$ и во втором интеграле для y: $a = 0,\ b = 1$ получим и для u: $a = 0,\ b = y$ получим:

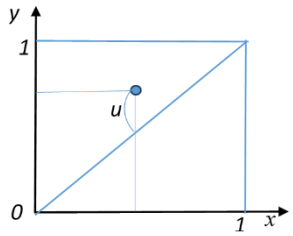


Рисунок 2.6 – Преобразование при выполнении замены переменных

$$\Delta c_{t,s}^{(i)} = \int_0^1 P_h(x)dx \int_0^x \left(\frac{x}{C + x\alpha_t^{(i)} + y\alpha_s^{(i)}} - \frac{y}{C + x\alpha_t^{(i)} + y\alpha_s^{(i)}} \right) P_h(y)dy +$$

$$+ \int_0^1 P_h(y)dy \int_0^y \left(\frac{x}{C + x\alpha_t^{(i)} + y\alpha_s^{(i)}} - \frac{y}{C + x\alpha_t^{(i)} + y\alpha_s^{(i)}} \right) P_h(x)dx$$

Теперь произведём замену переменных для первого интеграла $v = x$, $u = x - y$. Получим следующие пределы интегрирования:

$$0 < v < 1,\ 0 < u < v.$$

Якобиан преобразования имеет следующий вид:

$$\iint f(x,y)dxdy = \iint f(x(u,v), y(u,v)) \mid J \mid dudv,\ J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}$$

Выразив *x*, *y*

$$x = v,\ y = v - u$$

$$J = \begin{vmatrix} \frac{\partial v}{\partial u} & \frac{\partial v - u}{\partial u} \\ \frac{\partial v}{\partial v} & \frac{\partial v - u}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & -1 \\ 1 & 1 \end{vmatrix} = 1$$

Далее сделаем замену переменных для второго интеграла:

$$v = y,\ u = y - x$$

Получим следующие пределы интегрирования:

$$0 < v < 1,\ 0 < u < v$$

Выразим *x*, *y*:

$$x = v - u,\ y = v$$

Посчитаем Якобиан:

$$J = \begin{vmatrix} \frac{\partial v - u}{\partial u} & \frac{\partial v}{\partial u} \\ \frac{\partial v - u}{\partial v} & \frac{\partial v}{\partial v} \end{vmatrix} = \begin{vmatrix} -1 & 0 \\ 1 & 1 \end{vmatrix} = -1$$

С учетом этих выкладок получим:

$$\Delta c_{t,s}^{(i)} = \int_0^1 P_h(v)dv \int_0^v \left(\frac{u}{C + v\alpha_t^{(i)} + (v-u)\alpha_s^{(i)}} \right) P_h(v-u)du +$$

$$+ \int_0^1 P_h(v)dv \int_0^v \left(\frac{-u}{C + (v-u)\alpha_t^{(i)} + v\alpha_s^{(i)}} \right) P_h(v-u)u$$

Объединим интегралы:

$$\Delta c_{t,s}^{(i)} = \int_0^1 P_h(x)dx \int_0^x \left(\frac{x}{C + x\alpha_t^{(i)} + (x-u)\alpha_s^{(i)}} - \frac{x-u}{C + x\alpha_t^{(i)} + (x-u)\alpha_s^{(i)}} \right) P_h(x-u)du +$$

$$+ \int_0^1 P_h(y)dy \int_0^x \left(\frac{y-u}{C + (y-u)\alpha_t^{(i)} + y\alpha_s^{(i)}} - \frac{y}{C + (y-u)\alpha_t^{(i)} + y\alpha_s^{(i)}} \right) P_h(y-u)du =$$

$$= \int_0^1 P_h(x)dx \int_0^x \left(\frac{u}{C + x\alpha_t^{(i)} + (x-u)\alpha_s^{(i)}} \right) P_h(x-u)du + \int_0^1 P_h(y)dy \int_0^x \left(\frac{-u}{C + (y-u)\alpha_t^{(i)} + y\alpha_s^{(i)}} \right) P_h(y-u)du =$$

$$= \left[\int_0^1 \int_0^x \left(\frac{u}{C + x'\alpha_t^{(i)} + (x'-u)\alpha_s^{(i)}} \right) + \left(\frac{-u}{C + (x'-u)\alpha_t^{(i)} + x'\alpha_s^{(i)}} \right) \right] P_h(x')P_h(x'-u)dx'du .$$

В итоге можно прийти к следующему выражению:

$$\Delta c_{t,s}^{(i)} = \int_0^1 \int_0^v \left(\frac{u}{C + v\alpha_t^{(i)} + (v-u)\alpha_s^{(i)}} + \frac{-u}{C + (v-u)\alpha_t^{(i)} + v\alpha_s^{(i)}} \right) P_h(v)P_h(v-u)dvdu$$

Все остальные обоснования того, что подынтегральное выражение имеет отрицательный знак, ранее уже проведены для случая дискретного распределения.

Следствие 2. Как частный случай теперь выглядит классический метод DropKey [10-11], где для i -й строки выхода стохастическая составляющая

$$h_p^{(i)} = \{1,\ \Pr(\mathbf{h}^{(i)} = 1) = d_{key},\ h_p^{(i)} = \{0,\ \Pr(\mathbf{h}^{(i)} = 0) = 1 - d_{key}$$

осуществляет исключение весов внимания в соответствии с законом Бернулли. При этом при $\alpha_t^{(i)} > \alpha_s^{(i)}$ выполняется

$$c_t^{(i)} - c_s^{(i)} = (1 - d_{key})d_{key} \left[M_{a,h} \left[\frac{1}{\sum_{\substack{k=1 \\ k \neq t,s}}^n d_k^{(i)}\alpha_k^{(i)} + a_t^{(i)}} \right] - M_{a,h} \left[\frac{1}{\sum_{\substack{k=1 \\ k \neq t,s}}^n d_k^{(i)}\alpha_k^{(i)} + \alpha_s^{(i)}} \right] \right] < 0 .$$

Следствие 3. Пусть теперь при вычислении SA осуществляется добавление аддитивной составляющей , представляющей значение случайной величины $h^{(i)}$, вносящей независимое случайное воздействие на элементы строки матрицы внимания в позиции p перед выполнением нормировки в softmax. Выход в виде строки ответа на запрос для i -го пикселя (запроса) $q^{(i)}$ в данном случае будет иметь вид:

$$o^{(i)} = (o_1^{(i)},...,o_v^{(i)})^T = \tilde{\alpha}^{(i)} V = \tilde{\alpha}^{(i)} \left(v_1,...v_v\right),\ \tilde{\alpha}^{(i)} = (\tilde{\alpha}_1^{(i)},...,\tilde{\alpha}_n^{(i)}),\ i = \overline{1,n},$$

$$o_j^{(i)} = \tilde{\alpha}^{(i)} v_j = \sum_{p=1}^{n} \tilde{\alpha}_p^{(i)} v_j^{(p)},\ \ i = \overline{1,n},\ j = \overline{1,v}$$

$$\tilde{\alpha}_p^{(i)} = \frac{\exp\left(h_p^{(i)} + q^{(i)} k_p / \sqrt{d}\right)}{\sum_{p=1}^{n} \exp\left(h_p^{(i)} + q^{(i)} k_p / \sqrt{d}\right)} = \frac{\tilde{h}_p^{(i)} a_p^{(i)}}{\sum_{p=1}^{n} \tilde{h}_p^{(i)} a_p^{(i)}},\quad \tilde{h}_p^{(i)} = \exp\left(h_p^{(i)}\right),\ p = \overline{1,n}.$$

Таким образом, мы видим, что по форме полученное воздействие эквивалентно ранее рассмотренному мультипликативному и все доказательства возможности его использования для регуляризации роста весов внимания остаются в силе. Основное отличие состоит в том, что вносимая составляющая подвергается нелинейному преобразованию. При этом диапазон значений исходной случайной величины для преобразованной величины изменяется следующим образом: $\tilde{\mathbf{h}}^{(i)} \in [1,\ \mathrm{e}]$.

Обоснование целесообразности сглаживания. Для обоснования целесообразности подобной регуляризации весов внимания по аналогии с [93] рассмотрим оптимизационную задачу общего вида, которая решается численными методами в процессе выполнения нескольких итераций при обучении

$$\min_{a_p, v_p} \frac{1}{2} \left\| \sum_{p=1}^{n} a_p v_p - y \right\|^2,\quad \sum_{p=1}^{n} a_p = 1,\ \ a_p > 0, p = \overline{1,n}, \tag{2.10}$$

где $a_p \in \mathbf{R}, v_p \in \mathbf{R}^v$ – обучаемые параметры; $y \in \mathbf{R}^v$ – целевое значение. Обучаемые параметры можно рассматривать как вес и значение внимания. Представим каждое значение v_p в виде разложения на два вектора

$$v_p = \beta_p e + \gamma_p, \;\; e = \frac{y}{\|y\|}, \;\; y = Ce, \;\; C = \|y\|, \;\; \gamma^T y = 0,$$

где $\beta_p > 0$ – скаляр; e – единичный вектор, коллинеарный y, а γ_p – вектор ортогональный y. Для введенных векторов выполняется:

$$\rho^2\left(v_p, y\right) = \left(v_p - y\right)^T \left(v_p - y\right) = \beta_p^2 + \gamma_p^T \gamma_p - 2\beta_p C + C^2 = \left(\beta_p - C\right)^2 + \gamma_p^T \gamma_p, \;\; v_p^T y = \beta_p C,$$

При этом, если $v_p = y$, то $\rho^2\left(v_p, y\right) = 0$, $\beta_p = C$, $\gamma_p^T \gamma_p = 0$. При $v_p \neq y$ минимум евклидова расстояния $\rho\left(v_p, y\right) = \sqrt{\left[\gamma_p^T \gamma_p\right]}$ относительно β_p достигается при $\beta_p = C$. Из естественных соображений можно также предположить, что на начальных итерациях процесса обучения $\left|\beta_p^{(0)}\right| \ll C, \; p = \overline{1,n}$, а в процессе обучения все $\beta_p \to C$, $\gamma_p^T \gamma_p \to 0$. Возьмем частные производные целевой функции по обучаемым параметрам:

$$\frac{\partial L}{\partial a_p} = \left(\sum_{k=1}^{n} a_k v_k - y\right)^T v_p = \left(\sum_{k=1}^{n} a_k \left(\beta_k e + \gamma_k\right) - Ce\right)^T v_p = \sum_{k=1}^{n} \left(a_k \beta_k - C\right)\beta_p + \sum_{k=1}^{n} a_k \gamma_k^T \gamma_p,$$

$$\frac{\partial L}{\partial v_p} = a_p \left(\sum_{k=1}^{n} a_k v_k - y\right) = a_p \left(\sum_{k=1}^{n} a_k \left(\beta_k e + \gamma_k\right) - Ce\right) = a_p \left(\sum_{k=1}^{n} a_k \beta_k - C\right) e + a_p \sum_{k=1}^{n} a_k \gamma_k \; .$$

Теперь рассмотрим градиенты, рассчитываемые по обучаемым параметрам исходя из правил метода обратного распространения (Back Propagation**,** BP) ошибки

$$\frac{\partial L}{\partial v_p} = \frac{\partial L}{\partial \beta_p} e + \frac{\partial L}{\partial \gamma_p}, \quad \frac{\partial L}{\partial \beta_p} = a_p \left(\sum_{k=1}^{n} a_k \beta_k - C\right), \quad \frac{\partial L}{\partial \gamma_p} = a_p \left(\sum_{k=1}^{n} a_k \gamma_k\right), \qquad (2.11)$$

Нетрудно видеть, что они определяются компонентами векторного разложения (2.11). Отсюда следует, что, если вес $a_t > a_s$ и $\sum_{k=1}^{n} a_k \beta_k - C < 0$, то

$$\frac{\partial L}{\partial \beta_t} < 0,\ \frac{\partial L}{\partial \beta_s} < 0,\ \frac{\partial L}{\partial \beta_t} < \frac{\partial L}{\partial \beta_s},\ \left|\frac{\partial L}{\partial \beta_t}\right| > \left|\frac{\partial L}{\partial \beta_s}\right|. \quad (2.12)$$

Это означает, что скорость обновления для β_t будет больше, чем для β_s. Т.е. большее значение a будет способствовать увеличению β, а большее β будет способствовать дальнейшему увеличению a. Т.е. β_t будет быстрее расти и дальше в ходе обучения, стремясь к C. Аналогично для разницы градиентов относительно весов внимания

$$\frac{\partial L}{\partial a_t} - \frac{\partial L}{\partial a_s} = A(\beta_t, \beta_s) + B(\gamma_t, \gamma_s) = \left(\sum_{k=1}^{n} a_k \beta_k - C\right)(\beta_t - \beta_s) + \left(\sum_{k=1}^{n} a_k \gamma_k\right)^T (\gamma_t - \gamma_s) = \quad (2.13)$$

$$= \left(\sum_{k=1}^{n} a_k \beta_k - C\right)(\beta_t - \beta_s) + \left[\left(\sum_{k\neq t,s}^{n} a_k \gamma_k^T \gamma_t - \sum_{k\neq t,s}^{n} a_k \gamma_k^T \gamma_s\right) + \left(a_s \gamma_t^T \gamma_t - a_t \gamma_s^T \gamma_s\right)\right].$$

Анализируя (2.13) следует отметить, что первое слагаемое в указанных условиях будет меньше нуля и по модулю будет превосходить второе слагаемое. Последнее представляет скалярное произведение взвешенной суммы $\sum_{k=1}^{n} a_k \gamma_k$ векторов γ_k со случайным по знаку направлением, ортогональных вектору y, и разностного вектора $\gamma_t - \gamma_s$.

Как уже отмечалось, выполнение условия $A(\beta_t, \beta_s) < 0$ характерно на начальных этапах процесса обучения. Рассмотрим ситуацию инициализации обучаемых параметров. Поскольку все параметры инициализируются случайным образом, можно предположить, что начальное решение далеко от оптимального, то есть

$$a_1^{(0)} \simeq a_2^{(0)} \approx ... \approx a_n^{(0)} = \frac{1}{n},\ \left|\beta_p^{(0)}\right| \ll c,\ p = \overline{1,n}.$$

Тогда порядок первого и второго слагаемого в (2.13) будет определятся следующими соотношениями:

$$A(\beta_s,\beta_t)\simeq\left(\frac{1}{n}\sum_{p=1}^{n}\beta_p^{(0)}-C\right)\left(\beta_t^{(0)}-\beta_s^{(0)}\right)=\left(\bar{\beta}^{(0)}-C\right)\left(\beta_t^{(0)}-\beta_s^{(0)}\right)<0,$$

$$B(\gamma_s,\gamma_t)\simeq\frac{1}{n}\left(\gamma_s^T\gamma_s-\gamma_t^T\gamma_t\right)$$

и можно утверждать, что первый аддитивный член играет ведущую роль в (2.13). В итоге, выходные данные модели будут контролироваться несколькими разреженными блоками. И, наоборот, внесение искажений позволяет избежать неоправданного роста отдельных весов внимания и обеспечить более плавное распределение α.Следует отметить, что, хотя и представленное обоснование, в отличие от выполненного в предыдущем разделе, не является строгим доказательством, оно позволяет определить вероятные факторы, влияющие на целесообразность проведения сглаживающей регуляризации.

Оценка верхней границы степени сглаживания. Отдельный вопрос состоит в оценке степени сглаживания при регуляризации. Для оценки степени сглаживания при использовании базовой модели (2.5)-(2.8), т.е. конкретного соотношения $\Delta c_{t,s}^{(i)}<0,\quad c_t^{(i)}<c_s^{(i)}$ можно воспользоваться приведенными соотношениями для выполнения численных вычислений. Очевидно, что эти вычисления будут достаточно громоздкими и затратными по времени. Попробуем оценить верхнюю границу разности $\Delta c_{t,s}^{(i)}<0,\quad c_t^{(i)}<c_s^{(i)}$, опираясь на свойства суммируемых членов ряда в (6).

Для величин $E_{e,e-r}+E_{e-r,e}$ при произвольном $r=\overline{0,m}$ выполняются неравенства

$$E_{e,e-r}+E_{e-r,e}<0,$$

$$\left|E_{e,e-r}+E_{e-r,e}\right|<\left|\max\delta r\left[\frac{1}{\omega_{h,a}+I_e\alpha_t^{(i)}+I_{e-r}\alpha_s^{(i)}}P_t(I_e)P_s(I_{e-r})-\frac{1}{\omega_{h,a}+I_{e-r}\alpha_t^{(i)}+I_e\alpha_s^{(i)}}P_t(I_{e-r})P_s(I_e)\right]\right|,$$

$$\omega_{h,a}=\sum\nolimits_{k=1,k\neq t,s}^{n}h_k^{(i)}\alpha_k^{(i)}\ .$$

Величина разности в скобках определяется соотношением $I_e\alpha_t^{(i)}+I_{e-r}\alpha_s^{(i)}$ и $I_{e-r}\alpha_t^{(i)}+I_e\alpha_s^{(i)}$ при фиксированных $a_t^{(i)}>a_s^{(i)}$. Рассмотрим соотношение:

$$I_e\alpha_t^{(i)} + I_{e-r}\alpha_s^{(i)} - I_{e-r}\alpha_t^{(i)} - I_e\alpha_s^{(i)} = \delta r a_t^{(i)} + \delta r a_s^{(i)} < \max \delta r \left(a_t^{(i)} - a_s^{(i)} \right).$$

То есть мы видим, что максимум числителя и знаменателя достигается при одинаковых значениях, доставляющих этот максимум $I_e = 1,\ I_{e-1} = 0,\ \max \delta r = 1$. Отсюда получим, что

$$\left| E_{e,e-r} + E_{e-r,e} \right| < \left\| \left[\frac{1}{\omega_{h,a} + \alpha_t^{(i)}} P_t(I_e) P_s(I_{e-r}) - \frac{1}{\omega_{h,a} + I_e \alpha_s^{(i)}} P_t(I_{e-r}) P_s(I_e) \right] \right\|,$$

$$\omega_{h,a} = \sum_{\substack{k=1 \\ k \neq t,s}}^{n} h_k^{(i)} a_k^{(i)}.$$

Теперь, следуя введенному порядку суммирования в (6) можно оценить границу величины сглаживания

$$0 > c_t^{(i)} - c_s^{(i)} > \left[\frac{1}{\omega_{h,a} + \alpha_t^{(i)}} \sum_{e=0}^{m} \sum_{\substack{e=r \\ r>0}}^{m} P_t(I_e) P_s(I_{e-r}) - \frac{1}{\omega_{h,a} + \alpha_s^{(i)}} \sum_{\substack{e=r \\ r>0}}^{m} \sum_{e=0}^{m} P_t(I_{e-r}) P_s(I_e) \right]. \quad (2.14)$$

Суммы для вероятностей легко оценить следующим образом:

$$\sum_{e=0}^{m} \sum_{\substack{e=r \\ r>0}}^{m} P_t(I_e) P_s(I_{e-r}) = \frac{1}{2} \left(1 - \sum_{e=0}^{m} P_t(I_e) P_s(I_e) \right) = \frac{1}{2} \left(1 - \sum_{e=0}^{m} P^2(I_e) \right), \quad (2.15)$$

$$\sum_{\substack{e=0 \\ r>0}}^{m} \sum_{e=0}^{m} P_t(I_e) P_s(I_{e-r}) = \frac{1}{2} \left(1 - \sum_{e=0}^{m} P_t(I_e) P_s(I_e) \right) = \frac{1}{2} \left(1 - \sum_{e=0}^{m} P^2(I_e) \right).$$

Для дальнейшей оценки верхней границы докажем следующее неравенство:

$$\sum_{e=0}^{m} P^2(I_e) \le \frac{1}{m+1}, \quad \sum_{e=0}^{m} P(I_e) = 1 \ ,$$

т.е. минимум суммы квадратов вероятностей для полной группы событий достигается для равно вероятных событий $P(I_e) = 1/(m+1),\ \ e = \overline{0,m}$.

Доказательство проведем методом индукции. Пусть $m = 1$, тогда

$$P^2(0) + P(1)^2 = P^2(0) + (1 - P(0))^2,$$

$$\frac{dU}{dP(0)} = 2P(0) - 2(1 - P(0)) = 4P(0) - 2 = 0, \ \ P(0) = \frac{1}{2}, \ \ P(1) = \frac{1}{2}.$$

Предположим, что утверждение выполняется для произвольного $m>1$, т.е.

$$\min\sum_{e=0}^{m}P^2(I_e)=\frac{1}{m+1},\ \ P(I_e)=\frac{1}{m+1},e=\overline{0,m}.$$

Рассмотрим сумму для $m'>m+1$ и выполним ее преобразование

$$\sum_{e=0}^{m+1}P^2(I_e)=\sum_{e=0}^{m}P^2(I_e)+P^2(I_{m+1}),\ \ \sum_{e=0}^{m+1}P(I_e)=1.$$

$$U=\sum_{e=0}^{m+1}P^2(I_e)=\sum_{e=0}^{m}\left(P(I_e)+\frac{P(I_{m+1})}{m+1}-\frac{P(I_{m+1})}{m+1}\right)^2+P^2(I_{m+1})=$$

$$=\sum_{e=0}^{m}\left(P(I_e)+\frac{P(I_{m+1})}{m+1}\right)^2-\sum_{e=0}^{m}2\left(P(I_e)+\frac{P(I_{m+1})}{m+1}\right)\frac{P(I_{m+1})}{m+1}+\frac{P^2(I_{m+1})}{m+1}+P^2(I_{m+1})=$$

$$=\sum_{e=0}^{m}\left(P(I_e)+\frac{P(I_{m+1})}{m+1}\right)^2-2\frac{P(I_{m+1})}{m+1}+\frac{P^2(I_{m+1})}{m+1}+P^2(I_{m+1})=U'+U''.$$

Минимум суммы в $U'=1/(m+1)$ достигается как ранее доказано при

$$P(I_e)+\frac{P(I_{m+1})}{m+1}=\frac{1}{m+1},\ P(I_e)=\frac{1-P(I_{m+1})}{m+1},e=\overline{0,m},$$

а минимум оставшейся части при подстановке решения уравнения

$$\frac{dU''}{dP(I_m)}=-\frac{2}{m+1}+\frac{2P(I_{m+1})}{m+1}+2P(I_{m+1})=0,$$

$$P(I_{m+1})+(m+1)P(I_{m+1})=1,$$

$$P(I_{m+1})=\frac{1}{m+2}.$$

Доказательство завершено. Таким образом, можно оценить

$$c_t^{(i)}-c_s^{(i)}<0,\ \ \left|c_t^{(i)}-c_s^{(i)}\right|<\left|\left(\frac{1}{\omega_{h,a}+\alpha_t^{(i)}}-\frac{1}{\omega_{h,a}+\alpha_s^{(i)}}\right)\right|\left(1-\frac{1}{m+1}\right).$$

Отсюда, в частности, следует, при $m=1,\ I_0=0,\ I_1=1$ суммы в (2.15) преобразуются к виду:

$$\frac{1}{2}\left(1-\sum_{e=0}^{m}P^2(I_e)\right)=\frac{1}{2}\left(1-P^2(0)-P(1)^2\right)=\frac{1}{2}\left(1-P^2(0)-1+2P(0)-P^2(0)\right)=P(0)(1-P(0)$$

Таким образом, в данном частном случае мы получаем выражения, соответствующие технологии DropKey. Полученные выражения (2.14), (2.15) для границы степени сглаживания удобно использовать, если необходимо избежать громоздких вычислений по основным формулам. Далее рассмотрим частные способы структурной регуляризации весов внимания, применямые на практике.

2.3.2. Регуляризация весов внимания путем внесения аддитивной стохастической составляющей

Пусть теперь при вычислении самовнимания осуществляется добавление аддитивной составляющей $h_p^{(i)} \geq 0$, которая представляет значение случайной величины $\mathbf{h}^{(i)}$, вносящей независимое случайное воздействие на элементы строки матрицы внимания матрицы внимания в позиции p перед выполнением нормировки в softmax. Пусть для определенности $\mathbf{h}^{(i)}$ является случайной величиной, заданной на интервале [0,1] с распределением $P\left(h\right)=P_h(h_p^{(i)}),\ p=\overline{1,n},\ i=\overline{1,n}$. Пусть также исходные веса внимания в двух позициях (веса до внесения аддитивной составляющей) имеют соотношение $\alpha_t^{(i)} > a_s^{(i)}$. Оценим в этих условиях соотношение условных математических ожиданий величин $h_t^{(i)}, h_s^{(i)}$, осуществляющих коррекцию исходных весов внимания основе следующей цепочки уравнений:

Выход в виде строки ответа на запрос для i -го патча (запроса) $q^{(i)}$ в данном случае будет иметь вид:

$$o^{(i)}=(o_1^{(i)},...,o_v^{(i)})^T=\alpha^{(i)}V=\alpha^{(i)}\left(v_1,...v_v\right),$$

$$\alpha^{(i)}=(\alpha_1^{(i)},...,\alpha_n^{(i)})=soft\max\left\{h_1^{(i)}+q^{(i)}k_1,....,h_n^{(i)}+q^{(i)}k_n\right\}\ i=\overline{1,n},$$

$$\tilde{\alpha}_p^{(i)} = \frac{\exp\left(h_p^{(i)} + q^{(i)}k_p / \sqrt{d}\right)}{\sum_{p=1}^{n}\exp\left(h_p^{(i)} + q^{(i)}k_p / \sqrt{d}\right)},\ p = \overline{1,n},$$

$$o_j^{(i)} = \tilde{\alpha}^{(i)}v_j = \sum_{p=1}^{n}\tilde{\alpha}_p^{(i)}v_j^{(p)},\ \ i = \overline{1,n},\ j = \overline{1,\nu}.$$

Здесь $o_j^{(i)}$, как и ранее выход одного патча в ответ на воздействие запроса $q^{(i)}$. Исходя из свойств экспоненциальной функции, для величин $\tilde{\alpha}_p^{(i)}$ имеет место соотношение:

$$\tilde{\alpha}_p^{(i)} = \frac{\exp\left(h_p^{(i)}\right)\exp\left(q^{(i)}k_p / \sqrt{d}\right)}{\sum_{p=1}^{n}\exp\left(h_p^{(i)}\right)\exp\left(q^{(i)}k_p / \sqrt{d}\right)} = \frac{\tilde{h}_p^{(i)}a_p^{(i)}}{\sum_{p=1}^{n}\tilde{h}_p^{(i)}a_p^{(i)}},\quad \tilde{h}_p^{(i)} = \exp\left(h_p^{(i)}\right). \qquad (2.16)$$

Таким образом, мы видим, что по форме полученное воздействие сводится к ранее рассмотренному мультипликативному и все доказательства возможности его использования для регуляризации роста весов внимания остаются в силе. Основное отличие состоит в том, что вносимая составляющая подвергается нелинейному преобразованию. При этом диапазон значений исходной случайной величины для преобразованной величины изменяется следующим образом: $\tilde{\mathbf{h}}^{(i)} \in [1,\ \mathrm{e}]$.

Представленные выше исходные соотношения (2.4) предполагают выполнение двойной нормировки, что затратно в вычислительном отношении. Первый раз выполняется нормировка в softmax. Затем повторно после перемножения для всех произведений должна выполняться нормировка $h_p^{(i)}\alpha_p^{(i)}, p = \overline{1,n}$. Такая двойная нормировка выглядит неэффективной с вычислительной точки зрения. Поэтому, учитывая эквивалентность внесения мультипликативной составляющей для выхода softmax и аддитивной составляющей до выполнения операции softmax, целесообразно использовать следующие преобразования:

$$\tilde{\alpha}_p^{(i)} = \frac{\exp\left(\tilde{h}_p^{(i)} + \varepsilon + q^{(i)}k_p / \sqrt{d}\right)}{\sum_{p=1}^{n}\exp\left(\tilde{h}_p^{(i)} + \varepsilon + q^{(i)}k_p / \sqrt{d}\right)} = \frac{\exp\left(\tilde{h}_p^{(i)} + \varepsilon\right)\exp(q^{(i)}k_p / \sqrt{d}\right)}{\sum_{p=1}^{n}\exp\left(\tilde{h}_p^{(i)} + \varepsilon\right)\exp(q^{(i)}k_p / \sqrt{d}\right)},\ p=\overline{1,n},$$

$$o_j^{(i)} = \tilde{\alpha}^{(i)}v_j = \sum_{p=1}^{n}\tilde{\alpha}_p^{(i)}v_j^{(p)},\ \ i=\overline{1,n},\ j=\overline{1,v},$$

где $\tilde{h}_p^{(i)} = \ln\left(h_p^{(i)} + \varepsilon\right)$, а случайная величина $\mathbf{h}^{(i)} \in [0,1]$, как и ранее, имеет дискретное распределение $\mathbf{h}^{(i)} \in \{I_0,...I_m\},\ I_e = e/m,\ e=\overline{0,m}$, $P_p\left(I_e\right) = P_h(h_p^{(i)} = I_e),\ p=\overline{1,n}$ или непрерывное распределение $P_h(x)$. Величина ε обеспечивает исключение возможности возникновения неопределенности при использовании нулевых значения аргумента в логарифме и при этом не должна существенно влиять на вычисление ненулевых значений $h_p^{(i)}$. Потому ее целесообразно установить как $\varepsilon = \exp(\theta)$, где θ – достаточно большое по модулю отрицательное число, например $\theta = -10^6$.

2.3.3. Регуляризация весов внимания путем использования оценки корреляционных связей между элементами изображения

Очевидно, что больше каждое скалярное произведение (степень схожести пикселей), тем больший вес на выходе softmax и тем больше вес значения, которое взвешивается в формуле (2.3). Для удобства проведения дальнейших выкладок переобозначим векторы патчей как:

$$x_i = (x_{i,1},...,x_{i,d}) = q^{(i)} = k^{(i)},\ \ i=\overline{1,n}.$$

Рассмотрим оптимальную в классе линейных оценку каждого пикселя относительно другого. Стоит заметить, что это не глобально оптимальная оценка, которая в общем случая является нелинейной. Согласно [94] оптимальная линейная оценка (ОЛО) может быть получена в виде

$$\tilde{x}_{i/j} = \tilde{x}_i(x_j) = \tilde{M}[x_i / x_j] = M[x_i] + M[x_i x_j^T](M[x_j x_j^T])^{-1}(x_j - M[x_j]), \qquad (2.17)$$

где $\tilde{M}[x_i / x_j]$ обозначает условное математическое ожидание в широком смысле.

Учтем, что пиксели центрированы $M[x_i]=M[x_j]=0$. Тогда оценку в (2.17) можно представить в виде:

$$\tilde{x}_{i/j}=\tilde{x}_i(x_j)=\tilde{M}[x_i/x_j]=M[x_i x_j^T](M[x_j x_j^T])^{-1}x_j=R_{i,j}R_{j,j}^{-1}x_j,$$

где $R_{i,j}$ – матрица взаимной ковариации векторов x_i и x_j, а $R_{j,j}$ матрица ковариаций x_j, соответственно.

Важное свойство линейных оценок: статистическая ортогональность ошибки и самой оценки:

$$M[\left(x_i-\tilde{x}_{i/j}\right)x_j^T]=M[\left(x_i x_j^T-R_{i,j}R_{j,j}^{-1}x_j x_j^T\right)=R_{i,j}-R_{i,j}R_{j,j}^{-1}R_{i,j}=0,$$

$$M[\left(x_i-\tilde{x}_{i/j}\right)\tilde{x}_{i,j}^T]=M[\left(x_i x_j^T R_{j,j}^{-1}R_{i,j}^T-R_{i,j}R_{j,j}^{-1}x_j x_j^T R_{j,j}^{-1}R_{i,j}^T\right)=R_{i,j}R_{j,j}^{-1}R_{i,j}^T-R_{i,j}R_{j,j}^{-1}R_{i,j}^T=0$$

Матрица ошибок такой оценки имеет в общем случае вид:

$$E_{i,j}=M[\left(x_i-\tilde{x}_{i,j}\right)\left(x_i-\tilde{x}_{i,j}\right)^T]=M[\left(x_i-M[x_i]+M[x_i]-\tilde{x}_{i,j}\right)\left(x_i-M[x_i]+M[x_i]+M[x_i]\tilde{x}_{i,j}\right)^T]=$$

$$=M[\left(x_i-M[x_i]\right)\left(x_i-M[x_i]\right)^T]+M[\left(x_i-\tilde{x}_{i,j}\right)\left(x_i-\tilde{x}_{i,j}\right)^T]=$$

$$=R_{i,i}-R_{i,j}R_{j,j}^{-1}R_{i,j}.$$

В качестве показателей степени адекватности оценки можно использовать аналоги показателей регрессионного анализа TSS, ESS, RSS. В этом случае можно записать

$$TSS=trM[\left(x_i-M[x_i]\right)\left(x_i-M[x_i]\right)^T]=M[\left(x_i-\tilde{x}_{i,j}+\tilde{x}_{i,j}+M[x_i]\right)\left(x_i-\tilde{x}_{i,j}+\tilde{x}_{i,j}+M[x_i]\right)^T]=$$

$$=trM[\left(x_i-\tilde{x}_{i,j}\right)\left(x_i-\tilde{x}_{i,j}\right)^T]+M[\left(\tilde{x}_{i,j}-M[x_i]\right)\left(\tilde{x}_{i,j},-M[x_i]\right)^T]= \qquad (2.18)$$

$$=trM[\left(x_i-\tilde{x}_{i,j}\right)\left(x_i-\tilde{x}_{i,j}\right)^T]+M[\tilde{x}_{i,j}\tilde{x}_{i,j}^T],$$

$$RSS=trM[\left(x_i-\tilde{x}_{i,j}\right)\left(x_i-\tilde{x}_{i,j}\right)^T]=trE_{i,j}=trR_{i,i}-trR_{i,j}R_{j,j}^{-1}R_{i,j}^T,$$

$$ESS=trM[\tilde{x}_{i,j}\tilde{x}_{i,j}^T]=trM[R_{i,j}R_{j,j}^{-1}x_j x_j^T R_{j,j}^{-1}R_{i,j}^T]=trR_{i,j}R_{j,j}^{-1}R_{i,j}^T.$$

где tr – обозначает след матрицы.

Из (2.18) следует, что чем меньше RSS, тем лучше работает линейная модель оценки. Также, чем меньше ESS тем хуже она работает, в данном случае линейная связь x_i и x_j, описываемая $R_{i,j}$ мала. Отношение

$$T_R = 1 - RSS / TSS = ESS / TSS .$$

может служить мерой адекватности линейной модели.

Использование подобных линейных оценок в ходе глубокого обучения требует определения матриц взаимной ковариации между отдельными пикселями. Это возможно сделать, если использовать данные из разных изображений одного минипакета. Тем не менее, такой подход представляется весьма затратным в вычислительном отношении. Поэтому предлагается использовать некоторую априорную функцию пространственной корреляции между пикселями, вид которой определяется исходя из свойств анализируемых изображений. В качестве наиболее простого варианта можно рассматривать функцию пространственной корреляции однородного случайного марковского поля, которая определяется следующим образом:

$$R(x, x', y, y') = \sigma^2 \exp(-\alpha_x |x - x'| + \alpha_y |y - y'|) .$$

$$R[n, m] = \sigma^2 b_x^{n_x} b_y^{m_y}, \quad b_x = \exp(-\alpha_x \Delta x) \ , \ b_y = \exp(-\alpha_y \Delta y),$$

где α_x α_y – величины, обратно пропорциональные радиусам корреляции поля вдоль осей OX и OY соответственно изображения; σ^2 – дисперсия случайного поля; x, x', y, y' – координаты пикселей; $n = |x - x'| / \Delta_x$, $m = |y - y'| / \Delta_y$ – относительные целочисленные координаты.

Пусть также $\alpha_y = \alpha_x = \alpha$. Тогда для коэффициента корреляции между двумя любыми пикселями можно использовать следующее приближение:

$$R_{i,j} = M[x_i x_j^T] \approx \sigma^2 r_{i,j} = \sigma^2 b^{n_x + m_y}, \ \ n_x = |i_x - j_x|, \ \ m_y = |i_y - j_y|,$$

где параметр b может быть изменяемым или обучаемым.

Матрицу значений (или степени $n_{xy} = n_x + n_y$) для всех пикселей можно рассчитать заранее $R_{i,j}$, $i = \overline{1,n}$, $j = \overline{1,n}$, при этом величину b можно сделать обучаемой.

С учетом этого, в рамках используемой модели для ООЛ можно записать следующие соотношения:

$$\tilde{x}_{i/j} = \tilde{x}_i(x_j) = \tilde{M}[x_i / x_j] = r_{i,j} x_j ,$$

$$TSS = \sigma^2 ,\ E_{i,j} = RSS = \sigma^2 - \sigma^2 r_{i,j}^2 ,\ ESS = \sigma^2 r_{i,j}^2 .$$

Следует также отметить, что аналогично

$$\tilde{x}_{j/i} = \tilde{x}_j(x_i) = \tilde{M}[x_j / x_i] = r_{j,i} x_i ,\quad r_{j,i} = r_{i,j} .$$

Отсюда предлагается использовать для аддитивной коррекции стандартной меры внимания величину относительной среднеквадратичной ошибки, оптимальной в классе линейных оценок $\tilde{x}_{i/j=} r_{i,j} x_j$ пикселя x_i относительно пикселя x_j, которая имеет вид:

$$\delta_{j/i} = \left(x_i - r_{i,j} x_j\right)^2 \Big/ \left(\sigma^2 - \sigma^2 r_{i,j}^2\right).$$

Величина $\delta_{j/i}$ обладает следующими свойствами.

1.Если $r_{i,j} \to 0$, то $\delta_{j/i} \to x_i^2 / \sigma^2$ т.е. для отдаленных пикселей величина $\delta_{j/i}$ вносит одинаковые коррективы в стандартную формулу внимания, пропорциональные дисперсии того пикселя, с которым сравниваются другие пиксели.

2.Если $r_{i,j} \to 1$, то $\delta_{j/i} \to 0$, т.е. для рядом стоящих пикселей величина $\delta_{j/i}$ не вносит свои коррективы в стандартную формулу внимания.

3.Если ошибка $x_i - r_{i,j} x_j$ велика и $r_{i,j} > 0$, т.е. линейная модель связи не описывает реальную ситуацию (например, если x_i или x_j находятся на границе перепада яркости), $\delta_{j/i}$ вносит существенные коррективы в стандартную формулу внимания.

4.Если $x_i = x_j$, то $\delta_{j/i} = x_i^2(1-r_{ij}) / \sigma^2(1+r_{i,j})$ вносит коррективы, увеличивающиеся по мере отдаления пикселей друг от друга (результирующее внимание увеличивается).

Чтобы избежать деления на ноль при $i = j$ или при близких к единице $r_{i,j}$ выполняется простая регуляризация:

$$\delta_{j/i} = \frac{\left(x_i - r_{i,j}x_j\right)^2 + \varepsilon}{\left(\sigma^2 - \sigma^2 r_{i,j}^2\right) + \varepsilon},$$

где ε – малое положительное значение, неспособное повлиять на результаты вычислений в других ситуациях. Исходя из этих соотношений, получен алгоритм построения матрицы $D : D_{ij} = \delta_{j/i}$.

Нейронная сеть SwinIR уже имеет локальный механизм внимания, в рамках которого коэффициент пространственной корреляции должен меняться несильно. Следовательно, параметр b можно было подобрать оптимальным так, чтобы $r_{i,j}$ отличались друг от друга не более чем в 3 раза.

Пусть размер локального окна для механизма внимания $N \times N$. Тогда $b^{2N} = 1/3$. При размере окна в 8 пикселей значение b примерно равно 0,934. Этот параметр решено было сделать случайной величиной с равномерным распределением от 0,90 до 0,95 и менять его для каждого слоя с механизмом внимания.

На рисунке 2.7 показано значение пространственных коэффициентов корреляции для значения $b = 0.948$.

Рисунок 2.7 – Тепловая карта пространственного коэффициента корреляции. Расчет приводится для пикселя в координате (0, 0) и для патча размером 8x8.

Для ускорения вычислений был применен следующий механизм матричного умножения:

$$D = (X - X^T \otimes R)^2 / N^2,$$

где R – заранее рассчитанная матрица коэффициентов пространственной корреляции для каждого из пикселей окна размерами $N^2 \times N^2$, X – матрица значений пикселей следующего вида:

$$\begin{pmatrix} x_1 & x_1 & .. & x_1 \\ x_2 & x_2 & .. & x_2 \\ .. & .. & .. & .. \\ x_{N^2} & x_{N^2} & .. & x_{N^2} \end{pmatrix}.$$

От формулы аддитивной коррекции выражение отличается только знаменателем $\left(\sigma^2 - \sigma^2 r_{i,j}^2\right)$, зато дает высокую скорость восстановления

изображений из-за уменьшения числа операций, связанных с расчетом обратной матрицы, что было показано в работе автора [95].

Нормирующий множитель N^2 нужен, чтобы матрица D была именно корректирующей составляющей, а не вносила доминирующий вклад в матрицу внимания.

2.4. Использование обучаемой матрицы масштабных коэффициентов для сглаживания весов внимания

Еще один подход к модернизации механизма SA, предлагаемый в данной работе, основан на введении раздельно обучаемой матрицы S для регулирования масштабных коэффициентов скалярных произведений, вычисляемых до преобразования softmax [96]. Обучение элементов этой матрицы и весов внимания должно осуществляться раздельно. Математически подобное преобразование можно представить путем операции поэлементного перемножения матриц следующим образом:

$$R_s = S \otimes \frac{1}{\sqrt{d}} QK^T,$$

Соответствующая матрица весов значений после выполнения softmax построчно имеет вид

$$A_s = \begin{pmatrix} \alpha_1^{(1)} & \cdots & \alpha_n^{(1)} \\ \vdots & \vdots & \vdots \\ \alpha_1^{(n)} & \cdots & \alpha_n^{(n)} \end{pmatrix}, \quad \alpha^{(i)} = (\alpha_1^{(i)},...,\alpha_n^{(i)}) = \text{soft}\max\left\{\frac{s_{i,1}q^{(i)}k_1}{\sqrt{d}},....,\frac{s_{1,n}q^{(i)}k_n}{\sqrt{d}}\right\}, i = \overline{1,n}.$$

Таким образом, мы видим, что введение подобной матрицы эквивалентно введению раздельно обучаемого коэффициента масштабирования индивидуально для каждой позиции скалярного произведения исходной матрицы внимания. В известных работах подобный подход реализован только путем использования обучаемой скалярной переменной, как общего для них масштабного коэффициента. Начальные значения элементов матрицы S естественно установить следующим образом: $S = (s_{i,j}),\ s_{i,j} = 1,\ \ i = \overline{1,n},\ j = \overline{1,n}$.

Будем также считать, что на начальных итерациях процесса обучения все $s_{i,j}$ будут больше нуля. С целью анализа свойств введенной обучаемой матрицы запишем выражение для компонент вектора градиента целевой функции $L(o,y)$ (o,y – обобщенные обозначения выхода сети и целевого вектора) используемой при обучении трансформера методом обратного распространения ошибки (ВР). Согласно правилам алгоритма ВР частная производная относительно любого из обучаемых параметров $s_{i,p}, p=\overline{1,n}$ может быть вычислена на основе выражений для компонентов строк матрицы внимания $\alpha^{(i)}=(\alpha_1^{(i)},...,\alpha_n^{(i)})$:

$$\frac{\partial L}{\partial s_{i,p}}=\frac{\partial L}{\partial a_p^{(i)}}\frac{\partial \alpha_p^{(i)}}{\partial s_{i,p}}=L'_{a,i,p}a'_{s,i,p},\ \ \frac{\partial \alpha_p^{(i)}}{\partial s_{i,p}}=\frac{q^{(i)}k_p}{\sqrt{d}}\ f(\tilde{u}_{i,p})(1-f(\tilde{u}_{i,p})=u_{i,p}w_{i,p}(1-w_{i,p}),$$

(2.19)

$$\tilde{u}_{i,p}=s_{i,p}u_{i,p}=\frac{s_{i,p}q^{(i)}k_p}{\sqrt{d}},\quad w_{i,p}=f(u_{i,p})=\frac{\exp\left(u_{i,p}\right)}{\sum_{p=1}^{n}\exp\left(u_{i,p}\right)}.$$

Частная производная $L'_{a,i,p}$ при реализации ВР обычно содержит цепочку произведений производных функций активации, весовых коэффициентов связей последующих слоев нейронной сети и ошибок, распространяемых в обратном направлении.

Анализ (2.19) показывает, что знак частной производной $a'_{s,i,p}$ определяется знаком исходного скалярного произведения $u_{i,p}$. При этом, даже если значения скалярного произведения $u_{i,p}$ вводят функцию softmax в область насыщения ($f(u)\underset{u\to-\infty}{\to}0$ или $f(u)\underset{u\to+\infty}{\to}1$), частная производная по $s_{i,p}$ принимает существенные для продолжения обучения значения, что должно способствовать выходу из области насыщения. Если вес внимания $a_p^{(i)}$ требуется уменьшить при положительных значениях аргумента $\tilde{u}_{i,p}$ в (2.19) , то наиболее вероятно, что производная $L'_{a,i,p}$ будет больше нуля, тогда знак и значение $a'_{s,i,p}$ (при

положительном $s_{i,p}$) будет стимулировать уменьшать значение этого веса и, напротив, если требуется увеличить $a_p^{(i)}$ при отрицательном аргументе знак и значение $a'_{s,i,p}$ будет стимулировать увеличение соответствующего веса. Отметим, что именно положительные значения исходного скалярного произведения $u_{i,p}$ способны формировать превалирующие значения $w_{i,p} = f(u_{i,p})$ после softmax. Это, очевидно, можно считать регулирующим действием для механизма внимания.

Пусть $u_{i,p} \gg \varepsilon \geq u_{i,t},\ t = \overline{1,n},\ t \neq p,\ u_{i,p} > 0,\ \varepsilon > 0$ т.е. имеет место существенное превалирование одного элемента внимания в строке над другими. Нетрудно видеть, что в этом случае в строке выхода фактически формируется отклик, в котором основной вклад вносит пиксель, подаваемый на вход (при self-attention), причем практически без изменения. Если, например, существенно превалирующие элементы расположены по диагонали матрицы внимания, то выполняемые преобразования в трансформерном слое вообще не изменяют входную последовательность пикселей, что делает применение такого слоя бесполезным.

Пусть теперь, для пары элементов в строке матрицы внимания имеет место соотношение $\tilde{u}_{i,p} > \tilde{u}_{i,s} \geq 0$ и $s_{i,p}, s_{i,s}$ близки друг другу (для первого шага равенство точное). Соответственно $w_{i,p} = f(\tilde{u}_{i,p}) > w_{i,s} = f(\tilde{u}_{i,s})$. Тогда для частных производных по $s_{i,p}, s_{i,s}$ выполняется:

$$\frac{\partial \alpha_p^{(i)}}{\partial s_{i,p}} > \frac{\partial \alpha_s^{(i)}}{\partial s_{i,s}}. \tag{2.20}$$

Для доказательства (2.20) учтем, что производная функции softmax $y(w) = w(1-w) = -(w-0,5)^2 + 0,25,\ w = f(u)$ имеет немонотонный характер (см. рисунок 2.8) и симметрична относительно оси $x = 0,5$.

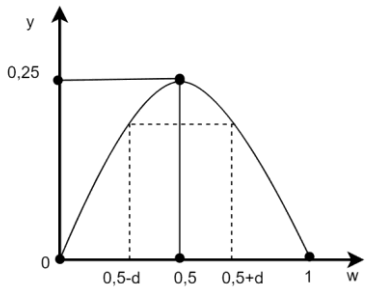


Рисунок 2.8 – Производная функции softmax

Пусть сначала $0 \le w_{i,s} < w_{i,p} < 0,5$, $w_{i,p} = f(\tilde{u}_{i,p})$, $w_{i,s} = f(\tilde{u}_{i,s})$. Тогда результат может быть получен непосредственно из графика функции $y(w)$ с учетом того, что ее производная $y'(w) = 1 - 2w > 0$, $w \in [0, 0,5)$ на этом участке положительна и монотонно возрастает:

$$\frac{\partial \alpha_p^{(i)}}{\partial s_{i,p}} = \frac{q^{(i)} k_p}{\sqrt{d}} f(\tilde{u}_{i,p})(1 - f(\tilde{u}_{i,p}) > \frac{q^{(i)} k_s}{\sqrt{d}} f(\tilde{u}_{i,s})(1 - f(\tilde{u}_{i,s}) = \frac{\partial \alpha_s^{(i)}}{\partial s_{i,s}},$$

$$0 \le f(\tilde{u}_{i,s}) < f(\tilde{u}_{i,p}) < 0,5$$

Пусть теперь $0,5 \le w_{i,p}$ или $w_{i,p} = 0,5 + d$, $0 \le d \le 0,5$ При этом обязательно, с учетом условия нормировки к единице, выполняемой softmax, значения $w_{i,s}$ должны лежать в диапазоне $0 \le w_{i,s} < 0,5 - d$. Тогда, как следует из рис.1, вследствие симметрии функции $y(w)$ относительно оси $x = 0,5$, и $y(w_{i,s})$ будет лежать в диапазоне значений $0 \le y(w_{i,s}) < y(0,5 - d) = y(w_{i,p})$. Тогда окончательно получим:

$$\frac{\partial \alpha_p^{(i)}}{\partial s_{i,p}} = \frac{q^{(i)} k_p}{\sqrt{d}} f\left(0,5 + d\right)\left(1 - f(0,5 + d)\right) = \frac{q^{(i)} k_p}{\sqrt{d}} f\left(0,5 - d\right)\left(1 - f(0,5 - d)\right) \ge \frac{q^{(i)} k_s}{\sqrt{d}} f(\tilde{u}_{i,s})(1 - \tilde{u}_{i,s}) = \frac{\partial \alpha_s^{(i)}}{\partial s_{i,s}} \cdot$$

Таким образом, можно утверждать, что применение раздельно обучаемой матрицы S в качестве мультипликативной составляющей при вычислении матрицы самовнимания в рамках стандартного механизма позволяет снижать воздействие возникающих аномалий в виде существенно превалирующих элементов самовнимания и включает дополнительные возможности регулирования весовых коэффициентов внимания.

Выводы по главе

1. Проведено исследование и визуализация карт внимания в моделях трансформерного типа для задач восстановления изображений. Выявлено, что механизм внимания чувствителен к типу и пространственным характеристикам помех на изображении. Не выявлено существенной зависимости между картами внимания и глобальными параметрами изображения. В тоже время, была выявлена зависимость между локальными участками канальных карт внимания и восстанавливаемым изображением. Что в целом доказывает обоснованность использования механизма внимания в моделях трансформерах, объясняя принцип его направленности на наиболее важные участки изображения при его восстановлении.

2.Произведено теоретическое обоснование и доказательство необходимости структурной регуляризации механизма внимания. В работе предложен и исследован подход к регуляризации механизма самовнимания в архитектурах трансформеров, направленный на повышение их эффективности и устойчивости. Теоретически обоснован метод внесения дискретной или непрерывной стохастической составляющей, используемой в механизме внимания, что позволяет сгладить неконтролируемую динамику роста весов в модулях SA в процессе обучения путем сглаживания существенно превалирующих весовых коэффициентов.

3. Рассмотрен подход к модификации механизма самовнимания, основанный на использовании обучаемой матрицы масштабирующих коэффициентов для матриц скалярных произведений. Теоретически показано, что применение подобной матрицы, которая вносится в схему обработки путем поэлементного перемножения с матрицами скалярных произведений в трансформерных блоках, способствует выходу активационной функции, применяемой в механизме внимания из области насыщения.

3. Синтез и анализ алгоритмов восстановления изображений на основе нейронных сетей-трансформеров

Теоретический анализ алгоритмов восстановление изображений с помощью ГНС трансформерного типа представлен в разделах выше. В текущей главе внимание будет уделено детальному описанию предложенных новых улучшенных архитектур подобных сетей, а также проведению сравнительных исследований предложенных решений с известными базовыми архитектурами.

При обучении сетей в качестве основной метрики количественной оценки качества восстановления изображений использовалось среднеквадратичное расстояние между пикселями истинного изображения и восстанавливаемого.

Помимо этого, для оценки итогового результата ВИ на тестовой выборке использовались другие известные метрики: отношение сигнал-шум *PSNR* и метрика структурного подобия *SSIM* для проверки качества изображений. Эти метрики являются общепринятыми в данной области, однако требуют наличия изображения эталона.

Метрика PSNR задавалась как:

$$PSNR(I) = 20\log_{10}\frac{\max I}{\sqrt{MSE}}, \tag{3.1}$$

где I исходное изображение; *MSE* – среднеквадратичная ошибка.

Основным преимуществом *SSIM* является то, что она точнее определяет различия между изображениями. Помимо этого, можно рассчитывать данную метрику по блокам изображения. Метрика принимает значения в диапазоне от 0 до 1 записывается следующим образом:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\delta_{xy} + c_2)}{(\mu_x{}^2 + \mu_y{}^2 + c_1)(\delta_x{}^2 + \delta_y{}^2 + c_2)}, \tag{3.2}$$

где μ_x, μ_y - математические ожидания изображений x, y, δ_x, δ_y - их стандартные отклонения, c_1, c_2 -поправочные коэффициенты

Многими исследователями было показано [88], что представленные выше метрики не всегда соответствуют характеру человеческого восприятия изображений. Поэтому решено было также использовать метрику Frechet Inception Distance (FID). Она основана на расстоянии Фреше между двумя распределениями признаков улучшенных и эталонных изображений. По сути, эта метрика показывает, насколько одно многомерное распределение похоже на другое. Для выделения признаков из изображений используется нейронная сеть InceptionV3, обученная на датасете ImageNet. В явном виде формулу можно записать следующим образом:

$$FID = \Sigma(\mu_1 - \mu_2)^2 + Tr(C_1 + C_2 - 2\sqrt{C_1 C_2}),$$

где μ_1 , μ_2 – вектора математических ожиданий по каждому из признаков эталонного и сгенерированного распределений; C_1, C_2 – матрицы ковариаций; Tr – след матрицы, а сумма берется по всем признакам изображений.

Из формулы можно сделать вывод, что чем меньше значение FID, тем ближе друг к другу распределения. Данная метрика уже зарекомендовала себя в различных задачах компьютерного зрения, поэтому решено было оценивать ею распределения, полученные из тестовой выборки, состоящей из 500 изображений для каждого далее используемого датасета.

Для проверки эффективности работы предложенных архитектур были выбраны следующие обучающие данные:

- ImageNet (50 тыс. изображений), с наложенными на него различными искажениями в виде аппликативных и аддитивных помех, сгенерированных авторами в соответствие с методикой, изложенной в [40];
- SIDD – набор данных, содержащий 30 000 зашумленных изображений из 10 сцен в различных условиях освещения, полученных с помощью пяти мобильных устройств, а также изображения их эталонов [97];
- Погодный датасет с 18 тыс. изображений, полученными в условиях снега, дождя, тумана, для которого в целях аугментации данных было решено

использовать библиотеку Albumentations [98], где уже реализованы алгоритмы генерации всевозможных погодных осадков.

Все изображения приводились к разрешению 256x256. Изображения из датасета SIDD могли иметь разрешение порядка 1024x1024, решено было улучшать их по частям. Предлагаемая архитектура может обработать каждую из частей изображения, а затем объединить их в единое, например, посредством билинейной интерполяции.

3.1. Предлагаемая архитектура трансформера с модифицированным механизмом канального внимания

Изначально в работе [41] выделено несколько основных проблем при использовании трансформеров для решения задач компьютерного зрения:

1. Длительный процесс обучения и сложная интерпретируемость результатов.
2. Квадратичная вычислительная сложность относительно числа пикселей изображения из-за использования механизмов внимания.
3. Необходимость использования большого набора данных для обучения.

На сегодняшний день эти проблемы остаются актуальными. Описанные в известных работах подходы отчасти решают проблему производительности и получения требуемого количества обучающих примеров, но не всегда позволяют использовать трансформеры в реальном времени на небольшом наборе данных в задачах машинного зрения.

В этом плане автор настоящей работы [99] предлагает модифицированный механизм внимания со сжатием канальной информации для уменьшения вычислительной сложности нейронной сети и увеличения ее быстродействия при сохранении относительно высоких показателей эффективности реализуемой обработки изображений.

Блоки нейронных сетей, в которых происходит сокращение канальной информации с последующим ее увеличением, именуется в литературе [16] как bottleneck. Отсюда, данную модификацию механизма внимания модели-

трансформер предлагается назвать как channel bottleneck self-attention mechanism, в дальнейшем CBSA-механизм. Функционально стандартный блок CBSA может быть представлен в виде, показанном на рисунке 3.1.

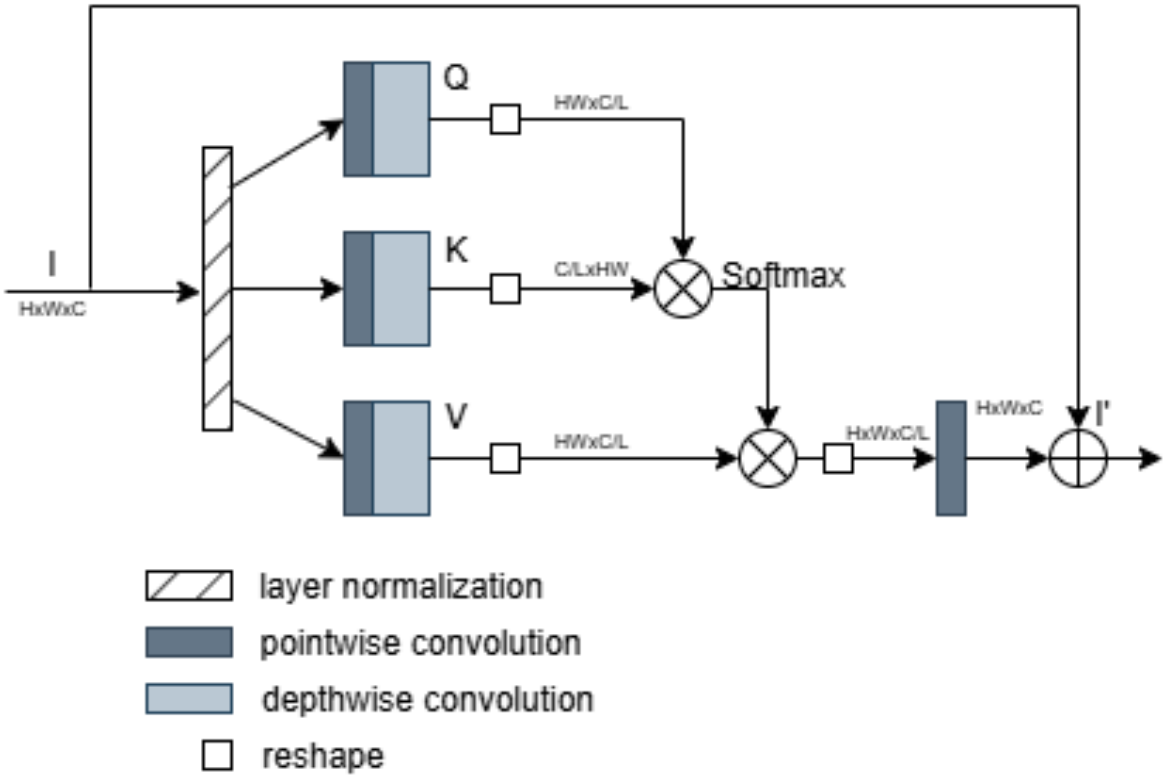


Рисунок 3.1. – Блок CBSA

Рассмотрим существо предлагаемого алгоритма сжатия канальной информации. Пусть есть нормализованный тензор I признаков размерамис, который подается на вход блока CBSA.

Этап 1: Канальное сжатие и разделение на Q, K, V. Применяется точечная свертка (с ядром 1×1) для извлечения объединенного тензора $\tilde{X}$:

$$\tilde{X} = \mathrm{Conv}_{1\times1}(I), \quad \tilde{X} \in \mathbf{R}^{H\times W\times 3C'}, C' = \frac{C}{L}.$$

Эта операция реализует как проекцию на Q, K, V, так и одновременно уменьшает число каналов в L раз. Далее применяется depthwise-свертка:

$$\tilde{X}^{dw} = \mathrm{DWConv}_{3\times3}(\tilde{X}),$$

где каждая из $3C'$ компонент обрабатывается независимым 3×3 фильтром.

После чего производится разбиение тензора по каналам на Q, K, V:

$$\tilde{X}^{dw} = [Q', K', V'], \quad Q', K', V' \in \mathbf{R}^{H \times W \times C'}$$

Этап 2: Преобразование в матричную форму. Для применения механизма SA, тензоры преобразуются в матрицы:

$$Q, K, V \in \mathbf{R}^{h \times d \times N}, \quad N = H \cdot W, \quad d = \frac{C'}{h},$$

где h – число голов внимания, а d – размерность канального подпространства для одной головы.

*Этап 3: Механизм самовнимания***.** В классической постановке механизм self-attention формализуется ранее приведенными соотношениями (2.1)-(2.3), но в формулах ниже он расписан для нескольких голов внимания, при этом переставлены местами канальная и пространственная размерности.

$$A = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right), \quad A \in \mathbf{R}^{h \times d \times d}$$

Полученные веса используются для взвешивания значений:

$$O = AV \in \mathbf{R}^{h \times d \times N}$$

Этап 4: Обратное преобразование и восстановление каналов. Происходит преобразования результата O в пространственный тензор: $Z' \in \mathbf{R}^{H \times W \times C'}$**,** после чего применяется pointwise-свертка, вместо классических полносвязных слоев с увеличением канальной информации в L раз:

$$Z = \text{Conv}_{1\times1}(Z') \in \mathbf{R}^{H \times W \times C}.$$

Изначально вычислительную сложность глобального механизма внимания для двух операций матричного умножения без учета голов внимания можно рассчитать следующим образом:

$$f(N) = O(N^2 C + N C^2).$$

В результате, с коэффициентом сжатия L итоговая вычислительная сложность становится равной:

$$f_R(N) = O(N^2 C / L + N C^2 / L)$$

и происходит уменьшение вычислительной сложности механизма самовнимания в L раз. Варьирование параметра L позволяет подбирать целесообразные его

значения, обеспечивающие уменьшение вычислительной сложности при незначительном снижении качества восстановления изображений. Данный подход позволяет эффективно учитывать, как пространственные, так и канальные зависимости, а также масштабировать архитектуру на входные изображения высокого разрешения.

В итоге, также уменьшается количество обучаемых параметров нейронной сети. В первую очередь за счет уменьшения количества фильтров для depthwise-сверток, так как до этого происходить сокращение числа каналов в L раз. Количество обучаемых параметров нейронной сети в зависимости от коэффициента сжатия представлено в таблице 3.1.

Таблица 3.1 – Количество обучаемых параметров нейронной сети

Степень сжатия	**Количество параметров**
L=1	26,111,668
L=2	22,379,476
L=4	20,513,380
L=8	19,580,332

Также необходимо было добавить нормализацию, чтобы увеличить сходимость и устойчивость нейронной сети. Полное отображение предлагаемой архитектуры нейронной сети представлена на рисунке 3.2. Вместо позиционного кодирования в ней использовался стандартный сверточный слой. Было показано в [36], что использование depthwise сверток вместо полносвязных слоев, следующих за механизмом внимания, значительно улучшает точность восстановления, поэтому этот же подход решено было применить в данной работе.

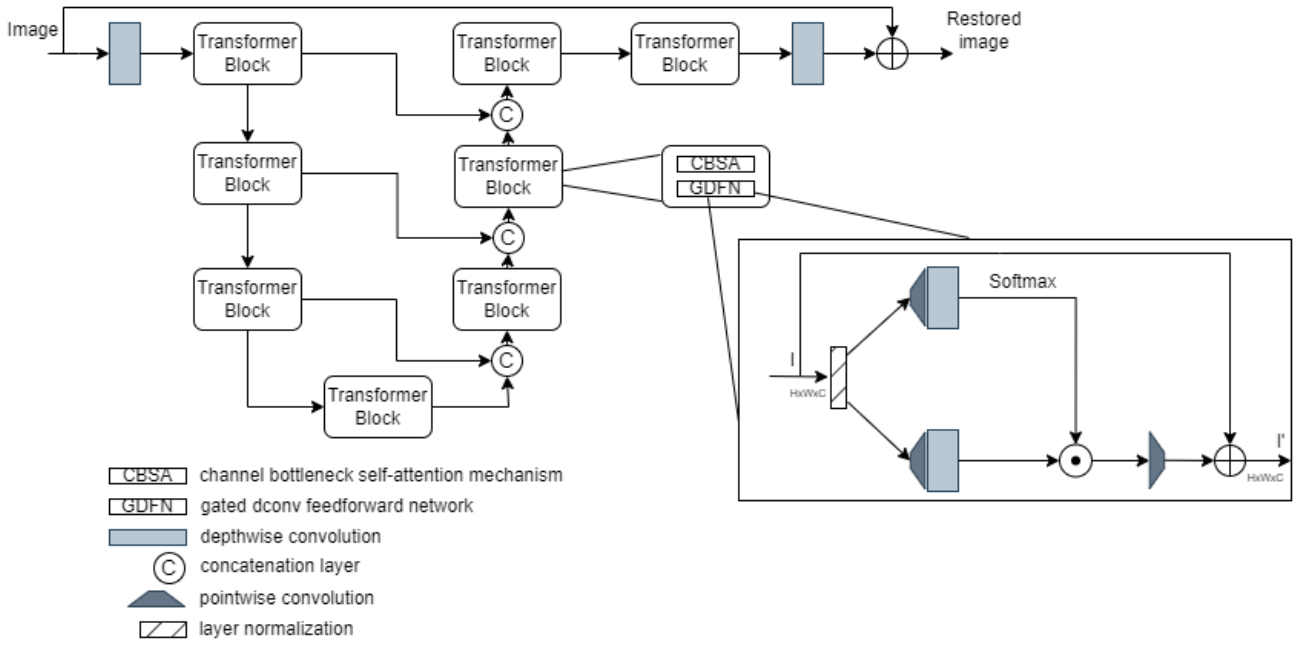


Рисунок 3.2 – Предлагаемая архитектура модели-трансформер для канального внимания

В работе [19] показано, что использование разделяемых сверток на ранних слоях нейронной сети может привести к падению точности. Поэтому решено было также использовать на первых двух блоках модели-трансформера обычные свертки.

Между блоками трансформера включены остаточные связи, необходимые для увеличения сходимости нейронной сети при обучении и для устранения проблемы затухания градиента, что очень критично для архитектур с большим количеством слоев. Изображение, полученное на выходе нейронной сети, показанной на рисунке 3.2, складывается с изображением на входе. Это является общепринятой практикой для задач улучшения качества изображений. Нейронной сети проще выучить убирать искажения, нежели формировать само изображение.

Предложенная архитектура нейронной сети обучалась на 8 Гб видеопамяти в течении 8 часов. В качестве оптимизатора использовался Adam. В качестве функций потерь предложено использовать вместо *MSE Charbonnier Loss* с добавлением *Edge Loss*, отвечающей за четкость восстанавливаемых границ,

определенных с помощью оператора Лапласа. Предлагаемую функцию потерь можно представить в виде формулы:

$$Loss(X,Y) = chl(X,Y) + w \cdot chl(\Delta X, \Delta Y),\ chl(X,Y) = \sqrt{(X-Y)^2 + e^2},$$

где Δ – оператор Лапласа; X, Y – улучшенное и эталонное изображения; w – коэффициент значимости функции потерь *Edge Loss*; e – коэффициент робастности для *Charbonnier Loss*.

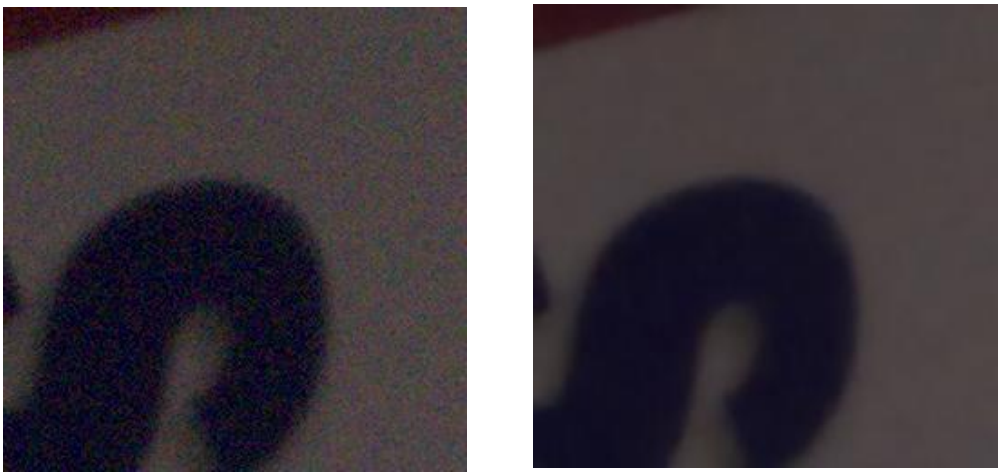
В ходе экспериментов были установлены оптимальные коэффициенты для функции потерь: $w = 1$ и $e = 10^{-3}$.

В работе была использована методика прогрессивного обучения, когда нейронная сеть учится улучшать качество изображений, начиная с маленького разрешения, а заканчивая самым большим. Таким образом, она становится адаптированной к различным разрешениям изображений, что является частым случаем в области обработки изображений. В таблице 3.2 показано результаты, полученные на датасете SIDD, при различных коэффициентах сжатия *L*.

Таблица 3.2 – Исследование качества изображения в зависимости от коэффициента сжатия *L* на датасете SIDD

SIDD	**PSNR**	**SSIM**	**FID**
L=1	39,23	0,97	47,42
L=2	39,14	0,97	48,17
L=4	39,15	0,97	49,46
L=8	37,1	0,95	50,36

На рисунке 3.3 показан пример зашумленного и улучшенного изображенияиз датасета SIDD.



а) б)

Рисунок 3.3 – Примеры изображений из датасета SIDD: а) – зашумленный патч изображения; б) – улучшенный патч изображения

Согласно метрикам оценки качества изображений, можно сделать вывод, что качество улучшенных изображений уменьшается с увеличением коэффициента сжатия *R.* Аналогичный вывод можно сделать на остальных датасетах. Результаты их исследований представлены в таблицах 3.3 и 3.4.

Таблица 3.3 – Исследование качества изображения в зависимости от коэффициента сжатия *L* на погодном датасете

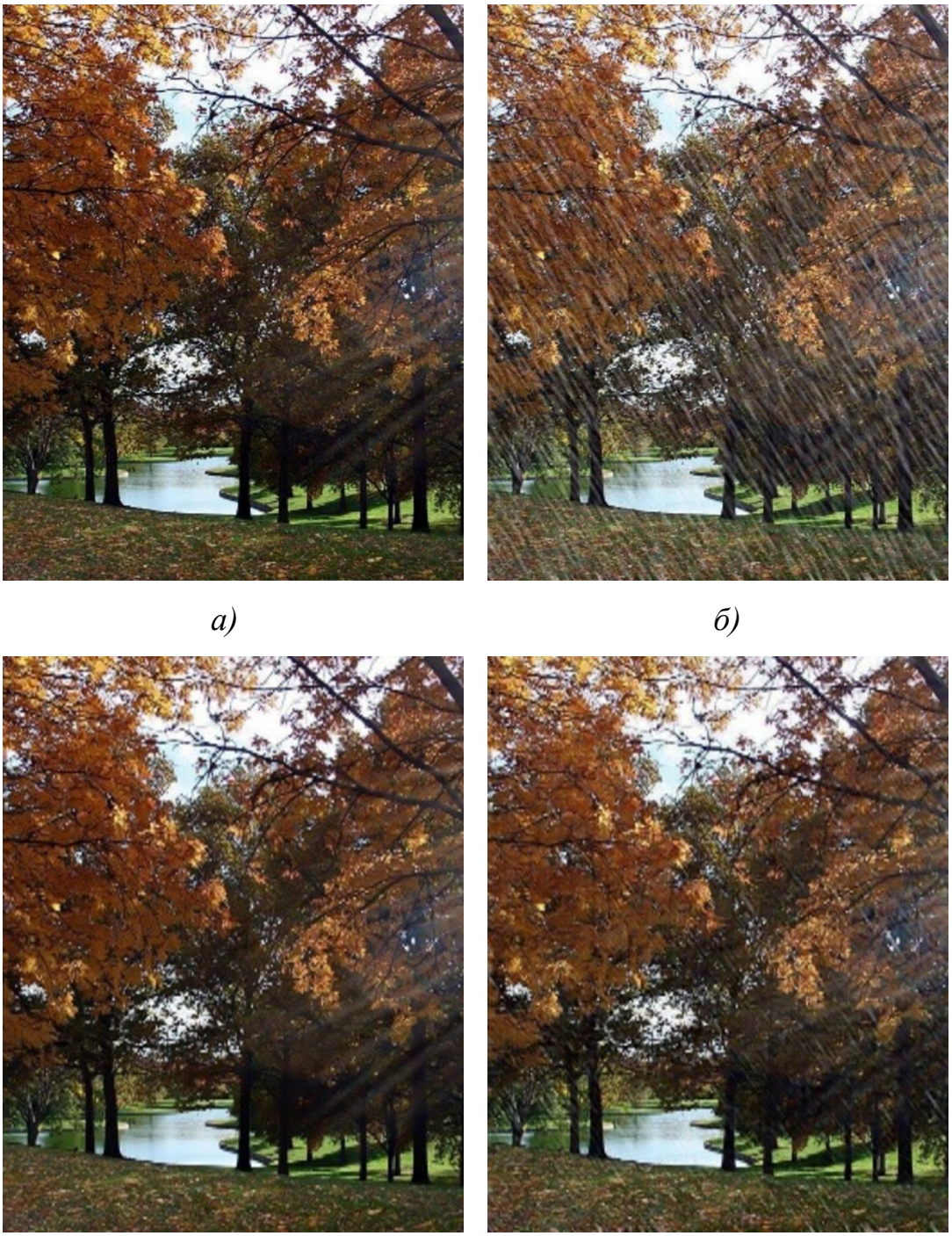
Погодный датасет	**PSNR**	**SSIM**	**FID**
L=1	26,6	0,93	74,42
L=2	26,6	0,92	74,75
L=4	26,5	0,92	75,46
L=8	26,0	0,90	76,46

Таблица 3.4 – Исследование качества изображения в зависимости от коэффициента сжатия *L* на зашумленном датасете ImageNet

Зашумлённый ImageNet	**PSNR**	**SSIM**	**FID**
L=1	38,32	0,75	102,06
L=2	38,20	0,74	102,73
L=4	38,14	0,74	103,18
L=8	38,01	0,74	103,44

В дальнейшем решено было остановиться на значении $L=2$, как на оптимальном выборе между качеством и быстродействием обработки

изображений. На рисунке 3.4 показано 4 изображения из датасета плохих погодных условий. Были специально выбраны изображения, содержащие текстурные части. Показано, что с увеличением коэффициента сжатия, нейронная сеть хуже справляется с восстановлением текстурной информации.



а) *б)*

в) *г)*

Рисунок. 3.4 Улучшенные фрагменты изображения из датасета SIDD: а) – исходное изображение; б) – изображение с каплями дождя; в) – улучшенное изображение при L=1; г) – улучшенное изображение при L =8

Для проведения сравненительного анализа были рассмотрены следующие архитектуры Restormer [36], SUNet [34], MIRNet [100]. MIRNet – сверточная нейронная сеть с механизмом внимания как для канальной, так и для пространственной информации. SUNet – модель-трансформер с локальным механизмом внимания с использованием принципа наложения пространственных окон. Restormer – гибридная модель-трансформер, использующая канальный механизм внимания. Все модели, обучались на представленных выше датасетах с оптимальной конфигурацией, предложенной их авторами.

В таблицах 3.5, 3.6, 3.7 приведены результаты исследования, описанных выше архитектур нейронных сетей, выделены наилучшие результаты для каждого из датасетов.

Таблица 3.5 – Сравнение на датасете SIDD

SIDD	**PSNR**	**SSIM**	**FID**
Restormer	**40,02**	0,96	**46,12**
SUNet	39,79	0,96	61,32
MIRNet	32,06	0,84	78,29
Авторский вариант (L=2)	39,14	**0,97**	48,17

Таблица 3.6 – Сравнение на датасете плохих погодных условий

Погодный датасет	**PSNR**	**SSIM**	**FID**
Restormer	**33,93**	0,91	83,15
SUNet	20,21	0,69	92,42
MIRNet	20,98	0,81	132,5
Авторский вариант (L=2)	26,6	**0,92**	**74,75**

Таблица 3.7 – Сравнение на зашумленном датасете ImageNet

Зашумленный ImageNet	**PSNR**	**SSIM**	**FID**
Restormer	36,11	**0,80**	**100,7**
SUNet	25,98	0,78	104,60
MIRNet	22,92	0,71	106,0
Авторский вариант (L=2)	**38,20**	0,74	102,73

Исходя из результатов, можно сделать вывод, что предлагаемая архитектура не уступает в точности Restormer на заявленных выше датасетах и метриках, а даже в некоторых случаях превосходит. Лучшие результаты выделены во всех таблицах. Однако имеет в 1,5 раз меньше параметров.

Отдельно проведены исследования абляции (возможность оптимизации за счет изменения количества используемых структурных элементов и некоторых гиперпараметров) предлагаемой архитектуры. В таблице 3.8 представлены результаты для датасета SIDD.

Таблица. 3.8 – Исследование предлагаемой архитектуры на оптимальность

SIDD	**PSNR**	**SSIM**	**FID**
Функция потерь Chabonier loss	39,01	0,96	48,42
Функция потерь MSE	38,79	0,96	51,33
Без прогрессивного обучения	39,06	0,93	50,27
Добавление двух блоков трансформеров	39,10	0,97	**48,11**
Удаление двух блоков трансформеров	37,56	0,92	53,14
Использование разделяемых сверток во всех блоках трансформеров	38,97	0,96	48,79
Авторский вариант (L=2)	**39,14**	**0,97**	48,17

Добавление двух блоков трансформеров: один – для понижения дискретизации, а второй – для повышения, незначительно улучшает результат. Как уже упоминалось выше, разделяемые свертки работают хуже, чем обычные в начальных блоках модели-трансформер. Важно отметить, что предлагаемая для обучения функция потерь Edge loss, показала себя лучше, чем обычная Chabonier loss. Также, можно отметить, что прогрессивное обучение положительно влияет на качество получаемых изображений.

3.2. Предлагаемая архитектура трансформера с модифицированным механизмом пространственного внимания

Базовой для решения задачи является работа [30], в которой для реализации локального механизма SA предложен swin-трансформер (shifted windows transformer) с использованием сканирующего окна и обучаемых параметров для кодирования блоков изображения. При этом размер блоков, в которых оценивается локальное внимание, составляет до 4x4 пикселей, а затем выполняется их иерархичное объединение. Это позволяет сделать постоянным число признаков на разных масштабах, поступающих на вход механизма SA. В настоящей работе, как прототип, используется сеть SwinIR [30], имеющая гибридную архитектуру, в которой можно выделить следующие особенности: поверхностное извлечение признаков с использованием сверточных слоев; глубокое извлечение признаков с использованием swin-трансформер; формирование матриц внимания в локальных блоках на основе скалярных произведений векторных представлений пикселей; высококачественная реконструкция изображений с помощью сверточных слоев.

В исходную архитектуру SwinIR, наряду с представленными выше модернизациями SA, были внесены определенные изменения, существо которых представлено на рисунке 3.5. Общая идея выполняемой в схеме рис. 3.5 обработки состоит в том, чтобы на основе изложенных в п.2.3 теоретических обоснований использовать для регуляризации и сглаживания весов пространственного внимания случайные величины, вычисляемые на основе значений выборочных статистик входных признаков, что позволяет получить для них наглядную физическую интерпретацию. Введение матрицы регулирования масштабированного скалярного произведения осуществляется по описанной выше схеме в разделе 2.

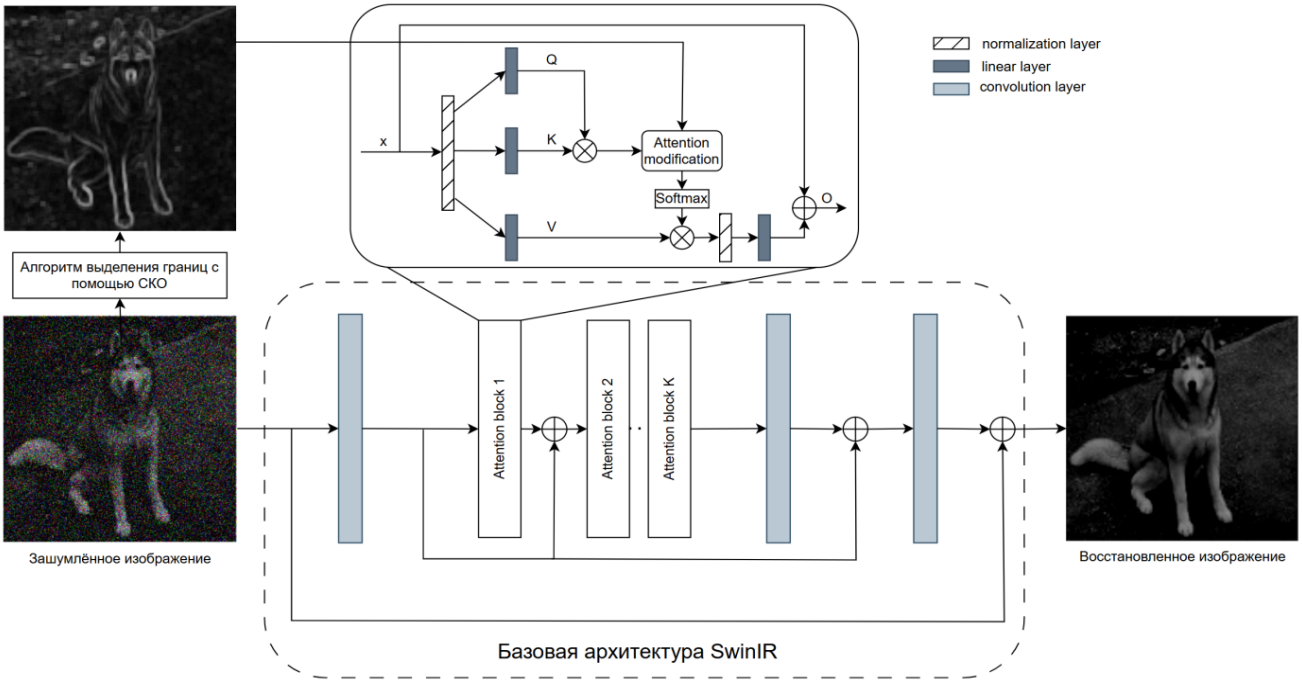


Рисунок 3.5 – Структурная схема архитектуры

Пусть имеется входное изображение I размерами $H \times W \times C$. В предлагаемом алгоритме сначала тензор I проходит через сверточный слой с ядром $k \times k$ с увеличением канальной размерности до F, при этом пространственная размерность сохраняется:

$$I' = \mathrm{Conv}_{k\times k}(I), \quad I' \in \mathbf{R}^{H\times W\times F}.$$

Далее пространство разбивается на непересекающиеся окна (patches) размером $N \times N$, в пределах которых будет применятся локальное самовнимание. Получается тензор размерности:

$$X \in \mathbf{R}^{P\times N^2\times F},$$

где $P = HW/N^2$ – количество патчей изображения.

Затем используется слой MLP для выполнения линейного погружения. На выходе получаем тензор размерами $P \times N^2 \times 3 \cdot F$, который затем разделяется на три одинаковых тензора K, Q, V.

Далее выполняется операция self-attention согласно формулам (2.1) - (2.3) с добавлением стохастической составляющей:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}+D\right)V,$$

где $Q = XW_Q$, $K = XW_K$, $V = XW_V$, с $W_Q,W_K,W_V \in \mathbf{R}^{F\times d}$; $Q,K,V \in \mathbf{R}^{P\times N^2\times d}$; $D \in \mathbf{R}^{P\times N^2\times N^2}$ – добавочная матрица со случайными элементами для структурной регуляризации.

Механизм SA в данном случае ограничен локальными областями $N \times N$, что снижает его вычислительную сложность в P раз. После применения self-attention каждый блок имеет форму:

$$O=\text{Attention}(Q,K,V) \in \mathbf{R}^{P\times N^2\times d}$$

Тензор O затем преобразуется обратно в пространственную форму со склейкой патчей:

$$Z' \in \mathbf{R}^{H\times W\times d}$$

После чего опять применяется MLP слой для восстановление канальной размерности F. В исследовании использовался частный случай, когда размерность d была равна F.

В архитектуре также присутствуют остаточные связи и используется механизм сдвига окон для обеспечения взаимодействия патчами, для которого выполняются те же самые описанные выше преобразования.

В качестве стохастических составляющих целесообразно использовать выборочные оценки локальных параметров входных признаков. Следовательно, все описанные выше теоретические обоснования для структурной регуляризации остаются действительны. Предлагается провести дисперсионный анализ изображения и построить тензор D для корректировки механизма внимания. Выполняются следующие действия:

1) Проводится отображение цветного изображения I в формат оттенков серого и нормировка следующим образом: $I_C = \left(I_{grey} - mean(I_{grey})\right) \Big/ std(I_{grey})$.

2) Выполняется скользящее сканирование изображения I_C окном размером $k \times k$ с единичным шагом и добавлением соответствующего reflect padding для сохранения размерности.

3) Оценивается выборочная дисперсия в скользящих окнах:

$$\sigma_{ij}^2 = \frac{1}{k^2 - 1} \sum_{a=i-\frac{k-1}{2}}^{i+\frac{k-1}{2}} \sum_{b=j-\frac{k-1}{2}}^{j+\frac{k-1}{2}} (I_{C,ab} - \overline{I_{C,ab}})^2 .$$

4) Для полученной матрицы $I_{std} = (\sigma_{ij})$ с целью устранения выбросов и уменьшения влияния шумов выполняется гауссовское размытие с маской 3×3. В итоге, получается матрица I_h, элементы которой используются далее для регуляризации внимания описанным выше способом.

Пример работы этой части алгоритма как раз показан на рисунке 3.5. Внизу исходное изображение, а вверху результат в виде матрицы I_h. Матрица фактически выделяет границы на изображении. Схожий результат можно получить дифференцированием, но вычисление дисперсии дало лучший результат при сильном зашумлении исходного изображения.

5) Полученная матрица I_h разбивается на блоки $N \times N$ и затем добавляется вспомогательная размерность. На выходе будет тензор I_{sp} размерностью $P \times N^2 \times 1$, где $P = HW / N^2$. Проводится формирование тензора для внесения в механизм SA $M_a = I_{sp} \cdot {I_{sp}}^T$ путем перемножения тензоров I_{sp} ($P \times N^2 \times 1$) и ${I_{sp}}^T$ ($P \times 1 \times N^2$) по последним двум измерениям.

Тензор M_a используется при реализации двух вариантов регуляризации. Первый вариант предполагает внесение аддитивной стохастической составляющей к матрице скалярных произведений в виде дисперсий окрестности исходных пикселей.

Второй – реализует адаптивный вариант DropKey со случайным удалением весов внимания, причем порог для удаления принимается с учетом информации об резких и текстурных областях, содержащихся в элементах M_a. Пусть

случайная величина R имеет равномерное распределение от 0 до 1. Тогда можно сформировать матрицу D следующим образом:

$$D_{c,ij} = \begin{cases} M_{a,cij}, & R \le r \,/\, M_{a,cij}, \\ -\inf, & R < r \,/\, M_{a,cij}, \end{cases}$$

где c – номер патча; i и j – пространственные индексы пикселей; r – задаваемая константа. Главное отличие от классического DropKey здесь в том, что порог не является фиксированным, а изменяется для каждого элемента матрицы внимания.

На рисунке 3.6 изображение карт внимания при данной модификации, демонстрирующее снижение структурной предвзятости механизма внимания.

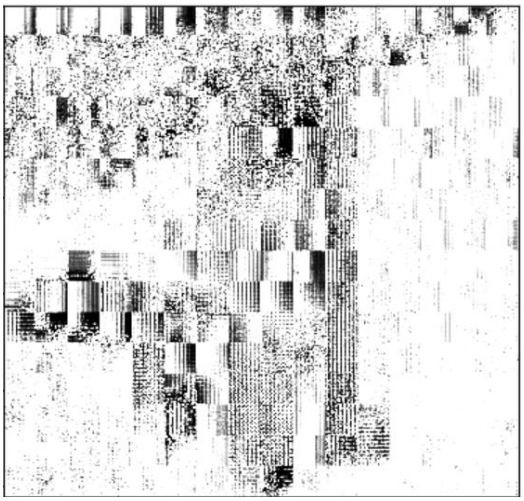


Рисунок 3.6 – Изображение карт внимания после структурной регуляризации (50 тысяч итераций при обучении)

Полученная в том или ином варианте матрица D служит аддитивной добавкой к матрице внимания. В таблице 3.8 представлены результаты восстановления изображений в зависимости от используемой модификации механизма внимания.

Все модификации являются частными случаями соотношений, доказанных выше для общего вида вносимой стохастической составляющей и обучаемой матрицы масштабированного внимания. Наилучший результат получен с помощью техники дообучения с отключением DropKey [93], поскольку данная техника нужна больше для регуляризации и на этапе тестирования в ней нет

нужды. Следует отметить, что в целом все предложенные схемы обработки показали преимущество по отношению к базовой архитектуре.

Таблица 3.8 – Результаты экспериментов для модификаций механизма SA

N	Использованные модификации SA	Подмножество ImageNet		Погодный датасет	
		PSNR	**SSIM**	**PSNR**	**SSIM**
1	Добавление D в виде выборочных дисперсий окрестности исходных пикселей	41,12	0,96	29,6	0,92
2	Адаптивный вариант DropKey	40,45	0,96	29,6	0,93
3	Стандартный DropKey [93]	40,20	0,93	28,0	0,90
4	Адаптивный вариант DropKey + 2000 итераций обучения с её отключением	**41,20**	**0,97**	**29,9**	**0,93**
5	Регулирование путем умножения на обучаемую матрицу S	40,17	0,94	29,2	0,93
6	Добавление D и регулирование путем умножения на S	41,04	0,97	29,5	0,91
7	Аддитивная коррекция на основе коэффициента корреляции	41,00	0,94	28,3	0,89
8	Исходная архитектура SWINIR [30]	38.18	0,90	25,0	0,85

Выводы по главе

1. Предложена схема модификации канального механизма внимания в моделях-трансформеров со сжатием канальной информации. Показано, каким образом используемый коэффициент сжатия L влияет на качество восстановленных изображений. Можно сделать вывод, что качество результатов нейронной сети падает незначительно, однако вычислительная сложность базовых блоков механизма внимания уменьшается в L раз. По большей части, это можно связать с тем, что сжатие положительно влияет на механизм внимания: имея меньше параметров, нейронная сети учится выделять только самое важное, меньше акцентируя внимание на незначимые признаки.

2. Рассмотрена модификация базовой архитектуры SwinIR, используемая в задачах восстановления изображений и реализующей локальное пространственное самовнимание в оконных блоках. Наибольшие усилия в это

части работы уделены реализации структурной регуляризации внимания за счет внесения по схеме, обоснованной в главе 2, стохастических составляющих в виде выборочных локальных оценок параметров изображений, в частности, выборочной дисперсии в локальных окрестностях пикселей.

3.На основе предложенной общей схемы обработки информации реализованы два способа регуляризации: аддитивное добавление матрицы выборочных дисперсий к скалярным произведениям и адаптивная модификация DropKey, учитывающая структуру изображения. В ряде вариантов также использовалась обучаемая матрица масштабирования. Результаты экспериментов на подмножестве ImageNet и погодном датасете показали, что все предложенные модификации превосходят базовую архитектуру SwinIR по метриках PSNR и по SSIM. Особенно эффективной оказалась стратегия использования DropKey только на этапе обучения.

3. Установлено, что все предложенные модификации показали увеличение эффективности работы глубоких нейронных сетей трансформерного типа. Модификации, реализованные в пространственном механизме внимания, смогли на практики подтвердить необходимость структурной регуляризации, теоретически обоснованной в разделе 2.

4. Синтез и анализ алгоритмов аугментации данных в задачах улучшения качества изображений. Структура программного комплекса для восстановления и аугментации изображений

В данной главе рассматриваются модели и алгоритмы генерации и аугментации изображений, направленные на расширение обучающих выборок и повышение устойчивости моделей компьютерного зрения к различным типам искажений. Представленные решения основаны как на эвристических методах обработки изображений, так и на современных архитектурах глубокого обучения в том числе генеративных и трансформерных моделях.

Приведенные ниже алгоритмы охватывают широкий спектр задач, включая:

- моделирование шумов и помех различной природы,
- реалистичный синтез погодных условий (осадки, туман, атмосферные искажения),
- частичную стилизацию изображений на основе нейросетевых блоков (в том числе AdaIN),
- а также генерацию синтетических обучающих данных с сохранением содержательной структуры сцены.

Все представленные в главе методы реализованы в виде программных компонентов в составе разработанного программного комплекса для восстановления и аугментации изображений. Структурное описание комплекса, архитектура и логика взаимодействия модулей представлены в пункте 4.4.

4.1. Алгоритмы внесения шумовых воздействий в обрабатываемые изображения

4.1.1. Эвристические алгоритмы внесения шумовых воздействий

Генерация дефекта на основе гауссовского распределения:

1) Случайным образом определялись размеры помехи h и w, формировалась матрица Np размерами $h \times w$.

2) Для генерации координат пикселя использовались случайные величины X и Y, имеющие нормальный закон распределения, с нулевым математическим ожиданием и дисперсиями $h/6$ и $w/6$ соответственно. Коэффициент $1/6$ был подобран, исходя из правила трех сигм для нормального распределения для того, чтобы подавляющее большинство значений лежало в отрезке от 0 до h и от 0 до w соответственно.

3) Формирование усеченного нормального распределения: ограничение сгенерированных значений соответствующими диапазонами от 0 до h и от 0 до w и округление их до ближайших целых x_i, y_i.

4) Увеличение значения яркости в соответствующем пикселе матрицы помехи:

$$Np[x_i, y_i] = Np[x_i, y_i] + 1.$$

5) Этапы 2 – 5 повторялись K раз. Параметр K был подобран экспериментальным путем $K = 3 \times h \times w$.

6) Нормировка значений матрицы Np: $Np[x_i, y_i] = Np[x_i, y_i] / max(Np)$.

7) Искажение объекта. Выбирался равновероятностно пиксель изображения, где будет находиться центр помехи. Далее происходило наложение искажения с сгенерированной помехой с исходным изображением.

Генерация аппликативных искажений (дефектов) методом наращивания локальных областей закрытия. Выбирались центры дефектов согласно пуассоновскому случайному процессу. Затем происходило их наращивание. Данный тип генерации был подробно описан в [101]. Авторы проводят исследование формы дефекта путем изменения определенных масок, меняющих топологию его наращивания. Изменяя вероятности искажения соседних пикселей относительно текущего заданием нормального и экспоненциальных распределений для наращивания области, можно контролировать прозрачность генерируемой помехи, что было показано в работе [102].

В данный подход была добавлена постобработка помех: применялся фильтр Гаусса с размером окна 3х3 для их размытия или, наоборот, использовался прием нерезкого маскирования для повышения их резкости и контрастности.

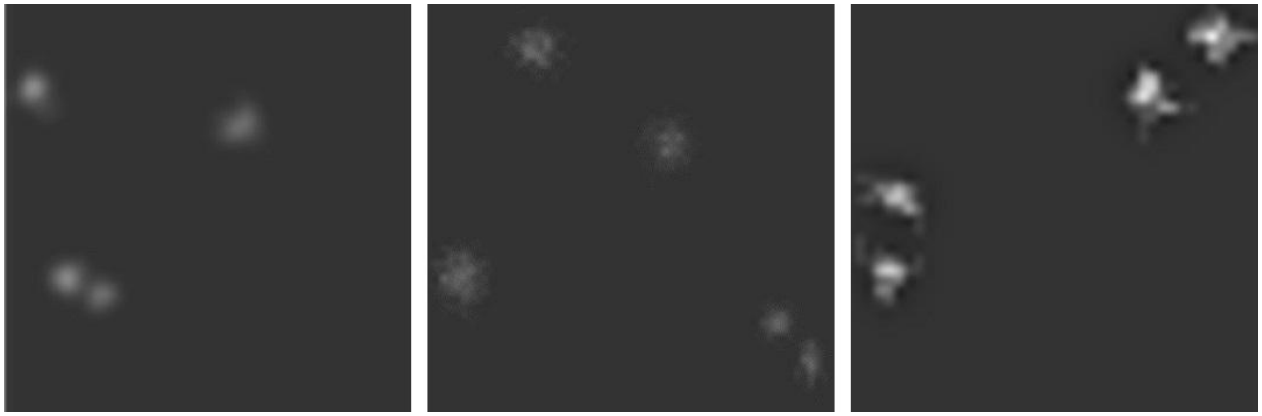
При использовании указанных выше двух способов не учитывалось наложение помех друг на друга. Рассматривались три случая искажения ими исходного изображения:

– помеха умножалась на коэффициент и складывалась с исходным изображением;

– помеха перемножалась с исходным изображением, происходило моделирование локальных мультипликативных помех;

– помеха замещала исходные пиксели изображения, происходило моделирование импульсных и аппликативных помех.

Примеры дефектов, сгенерированных способами 1 и 2, представлены на рисунке 4.1 в увеличенном размере и на черном фоне.



а) *б)* *в)*

Рисунок 4.1 – Сгенерированные дефекты на чёрном фоне: а) – генерация дефектов первым способом; б) и в) – генерация дефектов вторым способом с применением гауссовского фильтра и нерезкого маскирования соответственно

Генерация дефектов в виде краевой засветки изображения. Часто возникают помехи при неравномерном освещении исследуемого объекта. Предположим, что источник света находится сбоку от изображенных объектов. Согласно физическому закону обратных квадратов, интенсивность света обратно

пропорциональна квадрату расстояния от его источника. Для простоты будем считать, что световые волны, распространяющиеся от источника, имеют плоский фронт.

Засветку представим, как аддитивную помеху и будем добавлять ее только с одной стороны изображения согласно следующей формуле:

$$\hat{I}(x, y) = I(x, y) + h / (d + x)^2,$$

где d – предполагаемое расстояние до источника освещения; h – высота изображения; x, y – координаты пикселя исходного изображения I.

Указанная выше формула используется для генерации засветки от источника, находящегося справа от изображения. На рисунке 4.3 показано исходное изображение и зашумленное данным типом помехи.

Рисунок 4.2 – Генерация засветки внизу изображения

Генерация двоящихся границ объекта. Данный тип помех очень часто возникает при расфокусировки датчика камеры при съемке. Для его моделирования исходное изображение было размыто, чтобы не усиливать присутствующие на нем шумы и другие нежелательные артефакты. Затем был дважды применён прием нерезкого маскирования. После этого производилось смещение пикселей изображения на шаг d. На практике оно не задавалось больше 10. Далее полученное изображение накладывалось на исходное.

4.1.2. Частичная стилизация и блок AdaIN для ГНС сверточного типа

Стилизация изображений часто встречается в различных приложениях,

используется для получения живописных изображений с применением стилей художников. Однако в данной работе рассматривалось ее применение для аугментации данных. Примеры различных арт стилей, показывающие суть задачи стилизации изображений, представлены на рисунке 4.3. Здесь необходимо было контролировать стиль получаемых изображений для сохранения их реалистичности и, по возможности, сверять с эталонами. Задача усложнялась тем, что рассматривалось бесконечное количество стилей.

Рисунок 4.3 – Примеры различных стилей чайной кружки [103]

Из рисунка 4.3 виден основной эффект стилизации – форма объектов на изображении не искажается, меняется только стиль и цвет.

В настоящей работе на основе модификации архитектуры [104] был предложен новый способ искажения изображений, основанный на их частичной стилизации в реальном времени. В его основе лежит использование слоев AdaIN (adaptive instance normalization). Слои AdaIN похожи на обычную нормализацию экземпляров, но при этом не имеют обучаемых параметров, а вес и смещение вычисляются на основе изображения стиля. Формула нормализации представлена ниже:

$$AdaIN(x,y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right) + \mu(y),$$

где $\mu(x)$, $\sigma(x)$ – математическое ожидание и стандартное отклонение исходного изображения, а $\mu(y)$, $\sigma(y)$ – математическое ожидание и стандартное отклонение изображения стиля

Исходя из этого, решено было использовать собственный кодировщик, который обучался извлекать признаки непосредственно во время обучения

стилизации изображений. В качестве функции потерь была выбрана среднеквадратичная ошибка (MSE). При этом возникала проблема устойчивости при обучении данной сети, так как она не имела глобального ориентира и могла восстанавливать нереальные текстурные изображения. Ее удалось решить с использование дополнительной функции потерь – identity loss. Декодер помимо декодирования стилизованного изображения, также декодировал и исходное изображение.

Помимо этого, предложенная архитектура позволяет использовать параметр alpha, регулирующий степень стилизации изображений. Архитектура сети представлена рисунке 4.4. Она состоит из трех частей – кодировщика, декодировщика и слоёв AdaIN. Стрелки на схеме указывают, что AdaIN применяется для каждой остаточной связи исходной нейронной сети.

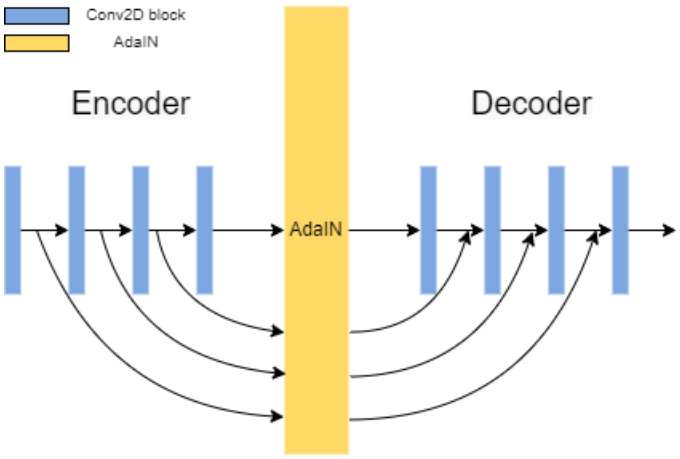


Рисунок 4.4 –Предложенная архитектура нейронной сети для стилизации изображений

Данная архитектура может использоваться для решения задач восстановления изображений, но в данном случае она использовалась для стилизации в двух вариантах:

1) Комбинирование стиля и исходного изображения при помощи AdaIN слоя. Использование собственного обучаемого кодировщика улучшает результат, но добавляет нестабильность при обучении. Для решения этой проблемы дважды хранились веса кодировщика. Первые из них изменялись стандартно, через алгоритм обратного распространения ошибки. А вторые веса менялись по

следующей формуле: $w_2 = aw_2 + (a - 1)w_1$. При тестировании в модель загружались только вторые веса, так как они были более стабильны при обучении.

2) Комбинирование классических алгоритмов стилизации и автоэнкодера. Использовались классические алгоритмы для повышения резкости, изменения яркости, контрастности и размытия изображений. Исходные изображения подавались на вход сети и на выходе сравнивались с уже примитивно стилизованными. Хотя данная архитектура имеет некоторую избыточность, считается, что нейронные сети более инвариантны к подаваемым на вход данным, следовательно, они будут действовать также эффективно на любых других изображениях. А вот для классических алгоритмов придется менять их параметры для того, чтобы эффективно провести стилизацию.

а) *б)* *в)*

Рисунок 4.5 – Частичная стилизация изображения. а) – исходное изображение, б) – накладываемый стиль, в) – стилизованное изображение.

Несмотря на полученный эффект для данной архитектуры необходимы дополнительные изображения-стили, а также отсутствие глобального ориентира не позволяет эффективно управлять стилем сгенерированного изображения. Эти проблемы были решены при помощи ГНС трансформерного типа и перекрестного механизма внимания, описанных ниже.

4.1.3. Эвристические алгоритмы генерации погодных осадков

Известные эвристические алгоритмы для генерации погодных осадков экспортированы из библиотеки Albumentations [98]. Библиотека Albumentations предоставляет набор специальных алгоритмов и трансформаций для генерации плохих погодных условий на изображениях. В частности, доступны классы RandomSnow (эффект снега), RandomRain (эффект дождя) и RandomFog (эффект

тумана). Эти трансформации накладывают на изображение искусственные осадки с помощью эвристических методов. Используются генерация шаблонов погодных осадков, геометрические фигуры, размытие и альфа-композиция, чтобы придать сцене реалистичный вид плохой погоды. Ниже описаны принципы работы каждой трансформации и используемые математические приемы.

RandomSnow создает эффект снежного покрова или снегопада, обесцвечивая часть изображения. Алгоритм работает в цветовой модели HLS, что позволяет изменять яркость без искажения оттенков. Задается порог яркости T, ниже которого пиксели считаются темными. Для таких пикселей яркость увеличивается по формуле:

$$L' = \begin{cases} \min(B \cdot L, L_{\max}), & L < T \\ L, & L \geq T \end{cases},$$

где L – исходная яркость пикселя; B – коэффициент яркости; $L_{\max}$ – максимум шкалы (например, 1.0 или 255).

Эффект проявляется в виде белых пятен на темных областях, имитирующие снег. Метод не отрисовывает отдельные снежинки, а реализует только снежный фильтр – эффект выбеливания сцены.

RandomRain накладывает на изображение полупрозрачные наклонные линии, имитирующие дождевые струи. Генерация включает следующие этапы:

1. Задание угла наклона θ, длины Δy и толщины линии.
2. Генерация координат начала каждой капли (x, y).
3. Построение линии от (x, y) до $(x + \Delta x, y + \Delta y)$, где $\Delta x = \tan(\theta) \cdot \Delta y$.

К линии применяется размытие, чтобы показать расфокусировку или засветку:

$$I_{\text{blur}}(x, y) = \frac{1}{49} \sum_{i=-3}^{3} \sum_{j=-3}^{3} I(x + i, y + j)$$

Также понижается яркость всей сцены:

$$L' = c \cdot L,$$

где $c \in (0,1]$ – коэффициент затемнения. В итоге получаются изображения с полупрозрачными дождевыми потоками, сниженной резкостью и яркостью.

RandomFog накладывает на изображение слой имитации тумана в виде полупрозрачных пятен. Туман формируется как набор светлых окружностей с радиусом, положением и прозрачностью. Центр изображения покрывается плотнее, края – менее заметны. Объединение тумана с изображением реализуется альфа-композицией:

$$I_{\text{out}}(x,y) = (1 - t(x,y)) \cdot I_{\text{orig}}(x,y) + t(x,y) \cdot C_{\text{fog}},$$

где $t(x,y)$ – совокупная непрозрачность тумана в точке (x,y); C_{fog} – цвет тумана (обычно белый). При перекрытии фигур $t(x,y)$ суммируется, до предела 1.0. Это создает плавный градиент плотности: от центра к краям. Размытие краев достигается перекрытием фигур разного размера и положением. Эффект напоминает наложение низкочастотной белой завесы, что снижает контраст и частично скрывает детали.

Эвристические трансформации RandomSnow, RandomRain и RandomFog реализуют погодные эффекты с помощью простых процедур: изменения яркости, наложения примитивов, размытия и альфа-композиции. Эти методы визуально правдоподобны и полезны для аугментации данных в задачах компьютерного зрения в условиях осадков и плохой видимости, но не пригодны для моделирования сложных атмосферных осадков и помех

4.2. Алгоритм синтеза изображений в условиях атмосферных осадков с помощью трансформера с перекрестным вниманием

В работе автора [105] предложена новая собственная модель WeatherTransformer, реализующая нейросетевую трансформерную архитектуру для условной генерации погодных эффектов на изображениях. Ключевая особенность WeatherTransformer состоит в использовании механизма перекрестного внимания (cross-attention), позволяющего вводить образец погодного эффекта (шаблон) и осуществлять формирование нового изображения

в соответствии с этим шаблоном. Под шаблоном в данной работе понимается любое изображение, содержащее на себе погодные осадки, не требующее отдельного выделения паттернов снега дождя и тумана на нем. Модель состоит из сверточных энкодера и декодера для иерархического извлечения и восстановления признаков с трансформерными блоками перекрёстного внимания. Для эффективного обучения введена составная функция потерь weatherLoss. На рисунке 4.6 представлена архитектура предложенной модели.

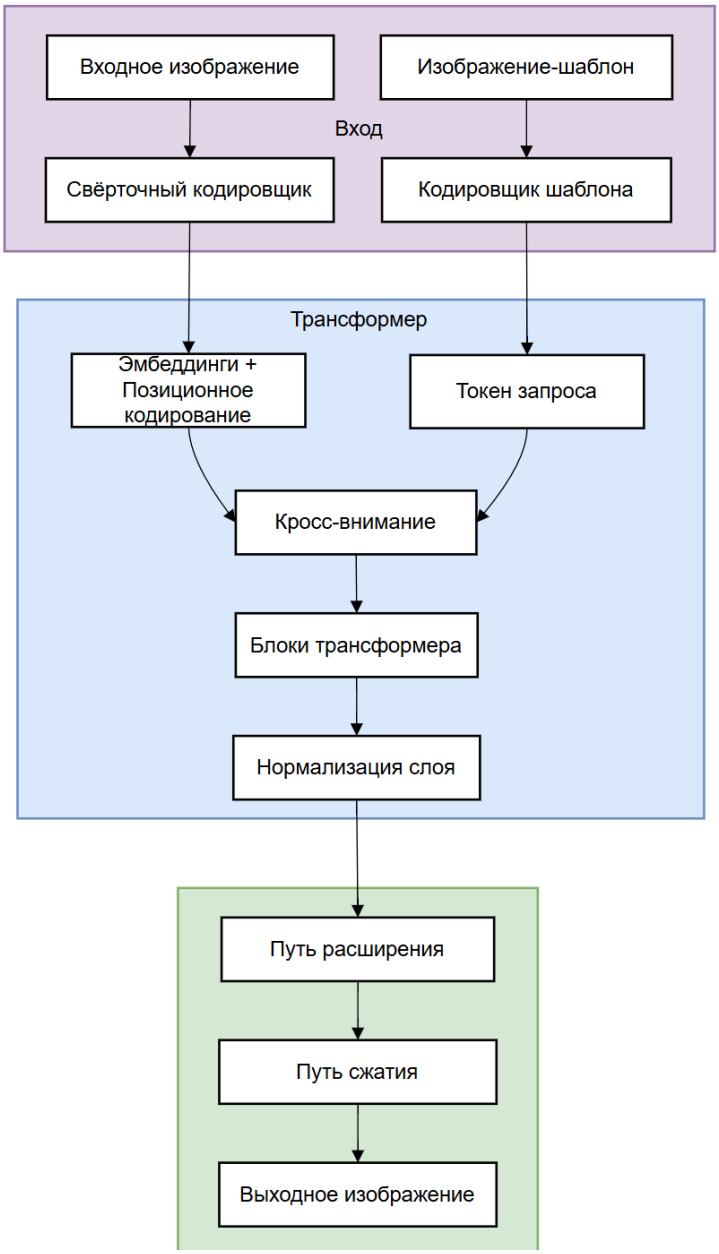


Рисунок 4.6 – Общая схема предлагаемой архитектуры.

Архитектура WeatherTransformer сочетает сильные стороны сверточных слоев (локальная инвариантность и иерархичность представлений) и трансформеров (глобальное внимание и адаптивность к контексту). Модель принимает на вход два изображения: исходное чистое изображение и изображение-шаблон с примером целевого погодного эффекта. На выходе формируется исходное изображение с включением наложенного эффекта погодного шаблона. Общая архитектура включает три основных блока: сверточный энкодер, трансформер с перекрестным вниманием и сверточный декодер. Кроме того, используется позиционное кодирование для сохранения пространственных связей между частями изображения.

Сверточный энкодер. Используется два одинаковых сверточных энкодера для извлечения признаковых карт из исходного изображения и шаблона. Энкодер имеет иерархическую многослойную структуру на основе сверточных блоков с поэтапным понижением разрешения изображений. Для входного RGB-изображения размером (256×256) энкодер выполняет последовательность следующих преобразований, показанных на рисунке 4.7.

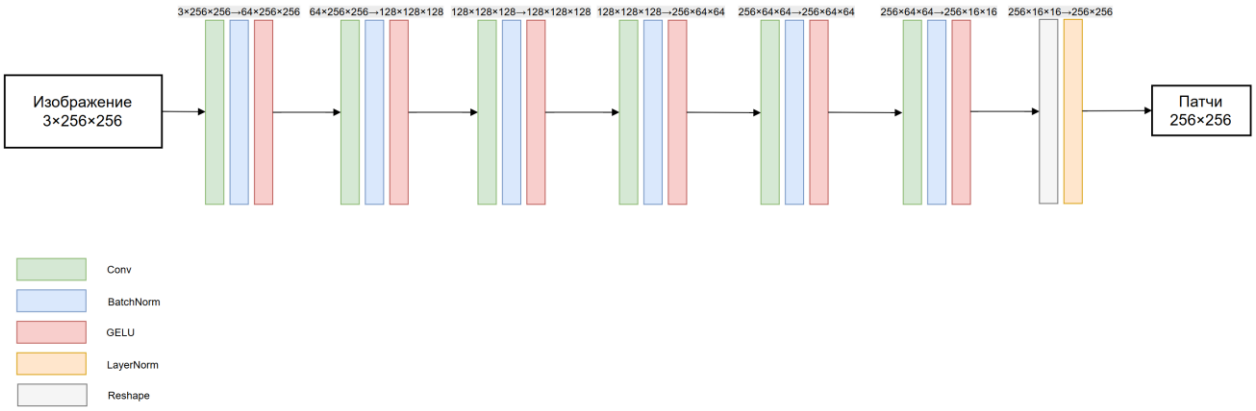


Рисунок 4.7 – Сверточный энкодер WeatherTrasformer.

На вход энкодера подается изображение $3 \times 256 \times 256$, затем увеличивается количество каналов до 64. Далее постепенно увеличивается количество каналов до 256 и уменьшаются пространственные размерности изображения до патчей 16×16.

Такой же энкодер используется для извлечения признаков из изображения-шаблона. Иерархическая структура энкодера позволяет выделять как локальные

детали, так и более крупномасштабные особенности сцены. В отличие от простого разбиения изображения на патчи (как в ViT [27]), перекрывающиеся свертки энкодера сохраняют пространственную структуру и обеспечивают инвариантность к мелким смещениям. Это создает признаковое представление, устойчивое к шуму и мелким вариациям, что важно при наложении погодных эффектов.

Перекрестное внимание и трансформерных блоках. После энкодера следует трансформерный модуль, состоящий из нескольких последовательных блоков трансформера с механизмом *cross-attention*. Реализуется взаимодействие между признаками шаблона (запросы) и признаками исходного изображения (ключи и значения). Перекрестное внимание позволяет шаблону определять важность и модифицировать определенные области изображения, перенося на них характерные черты погодного эффекта.

Каждый трансформерный блок реализует две стадии обработки: слоем перекрестного многоголового внимания и полносвязным слоем (FFN) для постобработки, причем для каждой стадии применяется остаточное соединение с нормализацией на входе (LayerNorm). Реализованный механизм внимания можно описать согласно формулам (2.1)-(2.3):

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + P\right)V,$$

где $Q \in \mathbf{R}^{m\times d}$ – матрица запросов (embedding патчей шаблона эффекта); $K,V \in \mathbf{R}^{n\times d}$ – матрицы ключей и значений (embedding патчей изображения); d – размерность ключей; а P – матрица относительных позиционных сдвигов пространственной размерностью, равной исходному изображению. За счет добавления матрицы P вычисляемые веса внимания учитывают взаимное расположение патча шаблона и патча изображения, что важно для корректного пространственного расположения эффекта. Результатом внимания является взвешенная комбинация значений V, соответствующая каждому запросу Q.

Чтобы понимать взаимосвязи между шаблоном и различными участками изображения, используется многоголовое внимание (multi-head attention). В работе использовалось $h=8$ голов: каждая голова вычисляет внимание для подпространства размерности d / h, после чего результаты объединяются для последующего линейного преобразования с помощью весовой матрицы W^{O}.

Сверточный декодер восстанавливает изображение с включенным эффектом из признакового представления, полученного трансформером. Декодер реализован как *прогрессивная* последовательность преобразований с увеличением разрешения изображения с помощью транспонированных сверток, чередующихся со сверточными блоками для уточнения признаков. Структура декодера показана на рисунке 4.8.

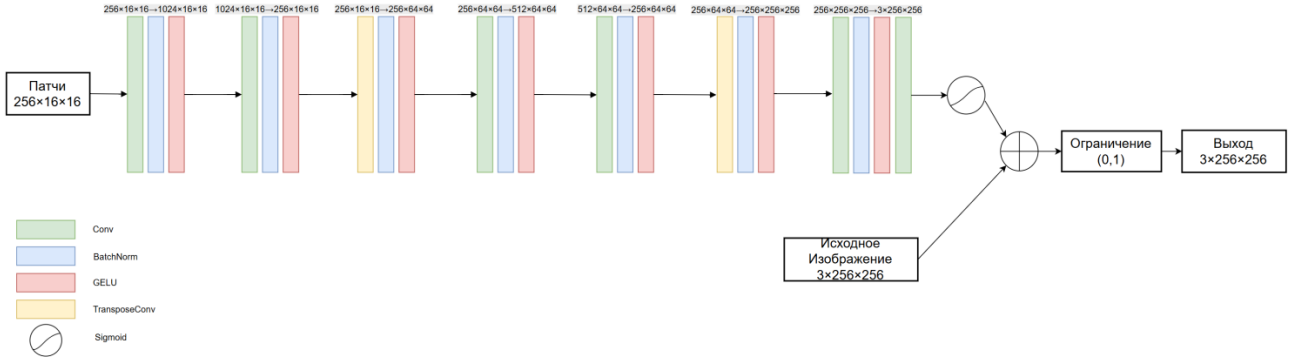


Рисунок 4.8. – Сверточный декодер WeatherTrasformer.

На вход декодера подаются патчи изображения $256 \times 16 \times 16$. Затем используется inverted bottleneck подход с первоначальным увеличением канальной информации в 2 раза для последующего сжатия и преобразования ее в пространственную.

Таким образом, архитектура WeatherTransformer объединяет энкодер, трансформер с перекрестным вниманием и прогрессивный декодер. Энкодер извлекает компактное описание сцены и шаблона, трансформер сопоставляет шаблон с содержимым сцены и передает паттерн эффекта на соответствующие его части, а декодер постепенно строит финальное изображение, добавляя погодный эффект к оригинальному изображению.

Следует также отметить, что, исходя из существа реализованного в данной модели подхода к процессу получения изображений с включением преднамеренных искажений, более уместно к нему применить термин синтез изображений, нежели генерация изображений.

Принцип обучения и взаимосвязь с обратной задачей – задачей восстановления изображений. Для обучения WeatherTransformer в каждом паттерне использовался набор из трех изображений: исходного чистого изображения, целевого изображения с наложенным эффектом (эталон, который модель должна получить) и изображения-шаблона, из которого извлекаются искажения и демонстрирующего характерные черты эффекта, связанного с наличием атмосферным порядком.

В качестве оптимизатора выбран AdamW, обеспечивающий стабильность при обучении трансформеров. Для автоматического подбора оптимальной скорости обучения в ходе обучения используется политика *OneCycleLR* [106]. Данная схема предполагает один цикл увеличения и затем плавного снижения скорости обучения. В начале обучения 10% итераций отводится на *разогрев* – скорость увеличивается от начального небольшого значения до максимального 10^{-4}, после чего в течение оставшихся 90% итераций она постепенно уменьшается по косинусной кривой до исходного уровня. Такая политика обучения (OneCycle) позволяет быстрее достичь высокой точности без ручного подбора скорости обучения, а также действует как регуляризатор, позволяя модели обучаться на повышенном градиентном шаге, что и показано в работе Smith и Topin [106] и это может улучшать обобщающую способность нейронной сети.

Помимо этого, используется ограничение градиента по норме для дополнительной устойчивости. Градиенты всех параметров обрезаются так, чтобы их евклидова норма не превышала 1.0. Это предотвращает *взрыв градиентов*, который может возникать при обучении трансформеров на высоких скоростях обучения, особенно на начальном этапе разогрева.

Функция потерь weatherLoss. Для включения погодного эффекта на исходное изображение необходимо учитывать несколько аспектов качества: контентное сходство (сохранение исходной сцены), структурное сходство (правильная передача контрастов и форм) и сходство эффекта на основе восприятия (визуальное соответствие наложенного эффекта желаемому стилю). Соответственно, была введена специальная функция потерь weatherLoss [105], которая сформирована как взвешенная сумма трех компонент, отвечающих за каждый из этих аспектов качества:

$$L_{total} = 0.7L_{content} + 0.2L_{struct} + 0.1L_{perc},$$

где коэффициенты 0.7/0.2/0.1 подобраны эмпирически, на основании предельных значений метрик и смысла так, чтобы контент имел бы наибольший приоритет, структура – меньший, а стилевой эффект – незначительный, но все равно учитывался бы на поздних стадиях обучения нейронной сети.

Для ошибки содержания $L_{content}$ использована L1 норма (MAE) между синтезированном изображением I_o и целевым изображением с эффектом I_t:

$$L_{content} = \frac{1}{3HW}\sum_{c=1}^{3}\sum_{x=1}^{H}\sum_{y=1}^{W}\left|I_o^{(c)}(x,y) - I_t^{(c)}(x,y)\right|.$$

Метрика MAE была выбрана вместо средней квадратичной ошибки, поскольку она менее чувствительна к выбросам и лучше сохраняет общее восприятие яркости/контраста.

Ошибка структурного сходства L_{struct} введена с целью сохранения ключевых структурных особенностей сцены. Она рассчитывается на основе индекса структурного подобия SSIM. SSIM – распространенная метрика оценки качества изображений, измеряющая сходство по яркости, контрасту и структуре между двумя изображениями. Перед вычислением метрики изображения пропускаются через фильтр Гаусса размера 11×11. Затем значение L_{struct} вычисляется между I_o и I_t следующим образом:

$$L_{struct} = 1 - SSIM\left(I_o, I_t\right).$$

Таким образом, максимальное структурное сходство (SSIM=1) даёт нулевую потерю, а любые отклонения от эталона снижают SSIM и увеличивают ошибку. Добавление компоненты L_{struct} в общую функцию потерь заставляет модель генерировать эффекты так, чтобы сохранялись отношения между соседними пикселями (локальные градиенты, текстуры) близкими к реальности.

Составляющая ошибка восприятия L_{perc} предназначена для оценки того, насколько визуально стиль синтезированного погодного эффекта соответствует заданному шаблону. В классических задачах стилизации изображений перцепционные (стилевые) потери вычисляются через сравнение высокоуровневых признаков, извлеченных предобученной сетью (например, активаций VGG).

В данной работе в целях увеличения скорости обучения выбран другой путь: прямое сравнение статистических характеристик синтезированного эффекта и эффекта шаблона в пространстве изображения. Пусть T – изображение-шаблон эффекта. Тогда для вычисления L_{perc} выполняются два шага преобразований.

1. *Совпадение паттерна***.** Из каждого изображения вычитается среднее значение по каждому каналу, и результат делится на стандартное отклонение по этому каналу: получаются нормализованные изображения I_o' и T'. Вычисляется среднеквадратичное отклонение (MSE) между I_o' и T' по всем пикселям:

$$L_{pattern} = \| I_o' - T' \|^2 .$$

Если, например, шаблон содержит характерную текстуру (узор бликов на каплях дождя или форму снежинок), то выход модели будет близок к шаблону, только если он воспроизвел этот узор независимо от глобальной разницы в интенсивности на изображениях.

2. *Совпадение статистики***.** Отдельно сравниваются первые два момента распределения пикселей выхода и шаблона. Вычисляются средние интенсивности μ_o, μ_T и стандартные отклонения σ_o, σ_T для каждого цветового канала выходного изображения и шаблона. Затем определяется ошибка:

$$L_{stats} = \sum_{c=1}^{3} \left[(\mu_o^{(c)} - \mu_T^{(c)})^2 + (\sigma_o^{(c)} - \sigma_T^{(c)})^2 \right].$$

Этот член штрафует расхождение по общей насыщенности эффекта (например, средней мутности тумана или общей контрастности снежинок) и по разбросу интенсивностей (контрасту текстуры эффекта) между выходом и шаблоном.

Таким образом, ошибку восприятия можно представить следующим образом:

$$L_{perc} = 0.5 L_{pattern} + 0.5 L_{stats}.$$

В итоге, L_{perc} принимает малые значения, когда и структура эффекта, и его основные статистические характеристики в синтезированном изображении соответствуют образцу.

Предложенная комбинированная функция L_{total} оптимизируется с помощью алгоритма стохастического градиентного спуска. В процессе обучения установлено, что каждая из составляющих играет свою роль: контентная L1 быстро снижает грубые пиксельные отличия, обеспечивая базовое сходство; структурная SSIM-подобная ошибка медленнее улучшает четкость и локальные детали; а ошибка восприятия, будучи наименьшей по весу, тонко настраивает визуальный стиль эффекта. Вместе они приводят к тому, что модель синтезирует изображения, практически неотличимые по содержанию от исходных и очень похожие по реализованному эффекту на указанный шаблон.

Для обучения предлагаемой модели был выбран датасет состоящий из 15 тысяч пар чистых и зашумлённых изображений со снегом, дождём и туманом. Для получения шаблона с помехой сначала её выделяли из зашумлённого изображения, затем она накладывалась на любое другое чистое изображение. При обучении гарантировалось, что зашумлённое изображение имеет такую же помеху, что и изображение-шаблон. При выделении помехи использовались, как эвристический алгоритм вычитания зашумлённого и чистого изображений, так и решение обратной задачи – восстановления зашумленных изображений с

помощью модели SwinIR. Последний алгоритм показал наилучший результат, но требовал дополнительных ресурсов на обучение нейронной сети для восстановления изображений. Такой подход не требует парных данных: чистых и зашумленных изображений. Достаточно иметь только зашумленные изображения, из которых можно получить восстановленные изображения и шаблоны. В таблице 4.1. показаны результаты синтеза изображений.

Таблица 4.1 – Синтез изображений для различных погодных условий

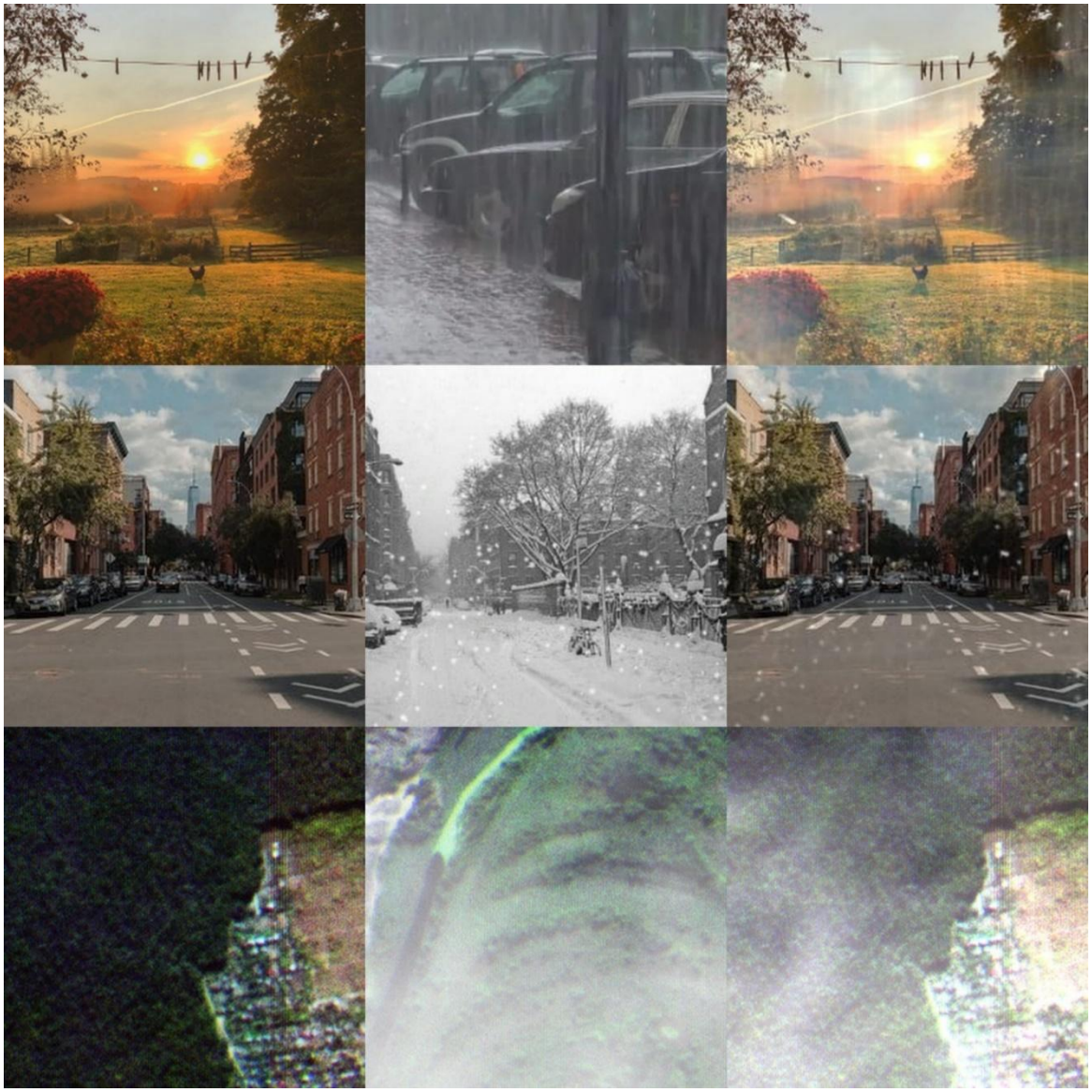
Тип погодного осадка	**PSNR**	**SSIM**	**FID**
Все погодные осадки	21.88	0.78	30.47
Только дождь	22.43	0.81	29.68
Только снег	22.5	0.81	29.70
Только туман	23.34	0.82	29.23

При проведении экспериментов был реализован универсальный вариант для синтеза изображений и использованием единой нейронной сети с добавлением шаблонов с любыми атмосферными осадками. Помимо этого, исследовались специализированные варианты с обучением отдельной нейронной сети для каждого вида осадков.

Из результатов в таблице 4.1 видно, что нейронная сеть показывает лучшие результаты, когда обучается на каждом типе погодных условий отдельно. Стоит также отметить, что нейронная сеть показала наилучшие результаты при синтезе изображений в условиях тумана, так как данный тип искажения значительно проще дождя и снежинок, из-за того, что не имеет чётко выраженную пространственную текстуру. На рисунке 4.9 показаны примеры синтезированных изображений.

А в таблице 4.2 проведены результаты исследования на оптимальность выбранной архитектуры, рассматривалось добавление нескольких дополнительных слоев, а также применение механизма самовнимания. Стоит

отметить, что выигрыш при добавлении дополнительных блоков оказался несущественным.



а) *б)* *в)*

Рисунок 4.9 – Примеры обработки изображений а) – исходные изображения, б) – изображения - шаблоны, в) –синтезированные изображение.

Таблица 4.2 – Проверка архитектуры на оптимальность

Модификация	**PSNR**	**SSIM**	**FID**
Дополнительный блок трансформер	21.97	0.78	30.22
Два дополнительных блока	**22.01**	**0.78**	**30.14**
Loss без perceptual	18.88	0.75	33.20
Self-attention	19.78	0.75	32.10

4.3. Применение алгоритмов аугментации данных в различных задачах компьютерного зрения

В целях проверки качества предложенных выше алгоритмов генерации, синтеза и стилизации изображений было решено использовать их для аугментации данных в различных задачах компьютерного зрения.

Сначала были рассмотрены алгоритмы наложения искусственных шумов и искажений. Исследовались две известные задачи машинного обучения: классификация и сегментация. Было проверено несколько способов комбинации исходных методов аугментации данных. В таблице 4.3 представлены результаты исследования. Для аугментации данных использовались описанные выше эвристические алгоритмы, частичная стилизация и генерация новых изображений при помощи BigGAN. В качестве архитектуры были выбраны ResNet для классификации и U-net для сегментации.

Таблица 4.3 – Аугментация данных на датасете кошек и собак

Алгоритм аугментации	Классификация на датасете кошек и собак **(accuracy)**	Сегментация на датасете кошек и собак **(IoU)**
Эвристические алгоритмы (Э. а.)	0.95	0.94
Э. а. + частичная стилизация	0.962	0.955
Э. а. + генерация изображений при помощи GAN	0.958	0.945

В задаче классификации решено было ограничиться использованием binary accuracy, так как дисбаланс классов не наблюдался. Исходя из полученных данных в таблице 4.3, можно отметить, что продвинутые техники аугментации данных могут внести большой вклад в повышение точности модели.

Помимо этого, было проведено сравнение качества восстановления изображений через задачу сегментации. Использовалась архитектура U-net, которая является одной из самых классических в области сегментации изображений. Особенно эффективно она показала себя при сегментации медицинских снимков [107].

Нейронная сеть обучалась на зашумленных и восстановленных изображениях. Далее сравнивалась точность ее работы, чтобы доказать факт того, что на восстановленных изображениях эффективность выше.

В качестве функции потерь использовался коэффициент Дайса. Он показал наилучшую эффективность по сравнению с MSE и IoU (intersection over union) функциями. Был взят датасет кошек и собак. Используя аргументированные данные, полученные при помощи частичной стилизации удалось улучшить результат сегментации согласно F1-меры на 1%.

В работе было проведено дополнительное исследование вопросов аугментации с использованием предложенной модели WeatherTrasformer в контексте применения синтезированных изображений для решения проблемы нехватки данных. В таблице 4.4 показаны полученные результаты для задачи восстановления изображений. Были использованы нейронные сети SwinIR и Restormer [36]. К обучающей выборке из 14 тысяч изображений было добавлено 1500 синтезированных изображений с их оригиналами, что позволило во всех случаях улучшить результат.

Таблица 4.4 – Аугментация данных на задаче восстановления изображений

Тип ГНС	Исходный датасет		Исходный датасет + синтезированный датасет	
	PSNR	**SSIM**	**PSNR**	**SSIM**
SwinIR	25.0	0.85	**25.5**	**0.86**
SwinIR с attention modification (Scale matrix)	29.9	0.93	**30.1**	**0.93**
Restormer	26.6	0.92	**27**	**0.923**

В табл. 4.5 и 4.6 представлены результаты, полученные на датасете автомобилей из Pascal VOC, для задач сегментации и детекции (обнаружения) объектов на изображениях. Для каждой задачи было использовано около 1500 размеченных данных. Было продемонстрировано, что в условиях существенной нехватки данных, аугментация с помощью модели WeatherTrasformer способна

значительно улучшить эффективность работы нейронной сети. В экспериментах к исходному датасету были добавлены изображения с включением различных атмосферных осадков на основе заранее выбранными шаблонами. В качестве архитектур были использованы устоявшиеся в данной области предобученные на датасете ImageNet модели: DeepLabV3+ с ResNet50 для сегментации и Faster R-CNN ResNet50 FPN V2 для обнаружения объектов.

Для сравнения результатов и оценки эффективности нейронных сетей были использованы стандартные метрики: Accuracy, IoU, Dice, F1-мера и mAP. Также была использована готовая библиотека Albumentations [98] для демонстрации важности качества синтезированных данных. Результаты показывают, что нейронные сети, обученные на датасете изображений, полученных на основе модели WeatherTransformer, показывают наилучший результат.

Таблица 4.5 – Аугментация данных в задаче сегментации

Задача	**IoU**	**Dice**
Сегментация Pascal VOC (DeepLabV3+ с ResNet50) без аугментации	0.82	0.85
Сегментация Pascal VOC (DeepLabV3+ с ResNet50) с Albumentations	0.83	0.87
Сегментация Pascal VOC (DeepLabV3+ с ResNet50) с WeatherTrasformer	**0.85**	**0.90**

Таблица 4.6. Аугментация данных на задаче обнаружения объектов

Задача	**F1-мера**	**mAP**
Детектирование Pascal VOC (Faster R-CNN ResNet50 FPN V2) без аугментации	0.87	0.84
Детектирование Pascal VOC (Faster R-CNN ResNet50 FPN V2) с WeatherTrasformer	**0.88**	**0.86**

4.4. Программный комплекс для восстановления и аугментации изображений

Как общий результат выполненных исследований и разработок автором предложен программный комплекс для восстановления и аугментации изображений (ПК ВИА). ПК ВИА представляет собой интегрированную

платформу для обработки изображений, объединяющую два функционально связанных направления обработки информации: восстановление изображений и генерацию шумов, искажений и погодных осадков для аугментации данных. Ключевая особенность ПК заключается в том, что эти задачи являются взаимообратимыми процессами – если алгоритм умеет качественно восстанавливать определенный тип искажений, то он может генерировать аналогичные искажения для обучения других моделей. На основные модули программного комплекса получено свидетельство о государственной регистрации программ для ЭВМ [108].

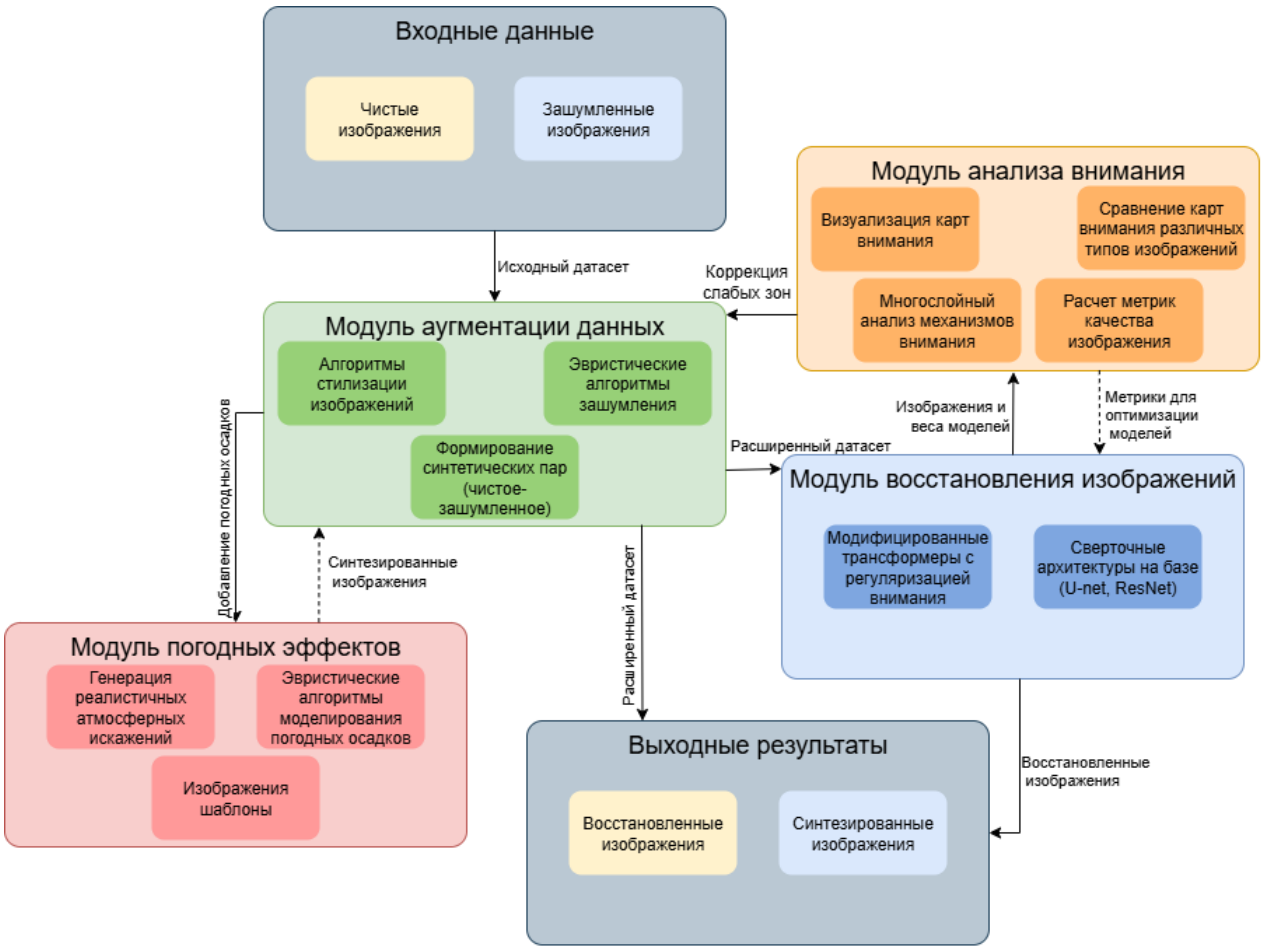


Рисунок 4.10 – Структурная схема предлагаемого программного комплекса

В рамках предложенной системы циклического обучения обосновано использование модульной архитектуры, обеспечивающей гибкость и адаптивность при работе с изображениями. Каждый модуль играет ключевую роль в соответствующих этапах цикла.

Модуль восстановления изображений включает как базовые сверточные архитектуры, так и модифицированные трансформеры с регуляризацией механизмов внимания, формируя основу первичного и повторного восстановления в итерациях обучения.

Модуль аугментации данных позволяет контролируемо расширять обучающую выборку за счет генерации различных искажений (включая зашумление и стилевой перенос и синтез изображений), что критически важно при формировании синтетических пар чистое-зашумленное изображений.

Модуль погодных эффектов реализует генерацию реалистичных атмосферных искажений с адаптивной интенсивностью, способствуя улучшению генеративной модели шума в обратной задаче.

Модуль анализа внимания используется на этапе проверки и отладки: визуализация карт внимания, многослойный анализ и сравнение различных типов внимания позволяют оценить интерпретируемость модели и уточнить ее слабые зоны для последующего дообучения.

Методика развертывания и применения программного комплекса (ПК ВИА) охватывает несколько этапов, соответствующих ключевым задачам в жизненном цикле обучения и применения нейросетевых моделей восстановления изображений.

Этап 1. Подготовка обучающей выборки. На первом этапе осуществляется формирование обучающего набора изображений, включающего как реальные, так и синтетически искаженные данные. Для генерации последних используется модуль аугментации данных, реализующий как стандартные эвристические искажения (зашумление, размытие, обесцвечивание), так и продвинутые методы переноса стиля и синтеза атмосферных условий. Особое внимание уделяется моделированию погодных явлений (дождь, снег, туман), реализованному через модуль погодных эффектов, в котором интенсивность и характер искажений адаптируются к контексту исходного изображения. Формируются пары изображений вида «искаженное – исходное», используемые для обучения моделей восстановления.

Этап 2. Обучение и настройка моделей восстановления. На данном этапе задействуется модуль восстановления изображений, включающий глубокие сверточные и трансформерные архитектуры, модифицированные с целью повышения устойчивости к шумам и улучшения обобщающей способности. В процессе итеративного обучения используются циклы генерации и восстановления: изображения из обучающей выборки проходят через алгоритмы аугментации и затем подаются в нейросетевую модель восстановления, которая учится восстанавливать исходный их вид. Результаты восстанавливаемых изображений сравниваются с эталонами с помощью стандартных метрик (PSNR, SSIM) и введенных автором визуально-интерпретируемых показателей.

Этап 3. Валидация и интерпретация моделей. На этапе валидации подключается модуль анализа внимания, предоставляющий визуализацию карт внимания и их динамику развития в процессе обучения. Это позволяет проводить исследование нейросетевой модели: выявлять области, на которые она фокусируется при восстановлении, и оценивать интерпретируемость механизма внимания. Использование этого модуля особенно важно при адаптации моделей под новые типы искажений, а также при выявлении «слепых зон» – участков изображений, где качество восстановления систематически оказывается неудовлетворительным.

Этап 4. Применение моделей и синтез новых искажений. После завершения обучения модели восстановления могут быть использованы не только на реальных искаженных изображений, но и в обратном режиме – в качестве генераторов новых синтетических данных. Такая взаимно-обратная архитектура реализует основной принцип ПК ВИА: если модель способна эффективно устранять определенный класс искажений, то она может быть использована для их воспроизведения с контролируемой интенсивностью и реалистичностью. Это позволяет динамически расширять обучающую выборку без необходимости ручной разметки.

Этап 5. Циклическое обучение и адаптация. ПК ВИА предполагает использование циклического обучения, при котором результаты анализа

внимания и эффективности восстановления на тестовых данных используются для дообучения или адаптации моделей. Например, при обнаружении ухудшения качества восстановления в специфических погодных условиях может быть инициирована повторная генерация обучающих данных с усилением соответствующего погодного эффекта. Таким образом, обеспечивается гибкость и адаптивность модели к изменениям условий эксплуатации.

Общий интерфейс ПК ВИА представлен на рисунке 4.11

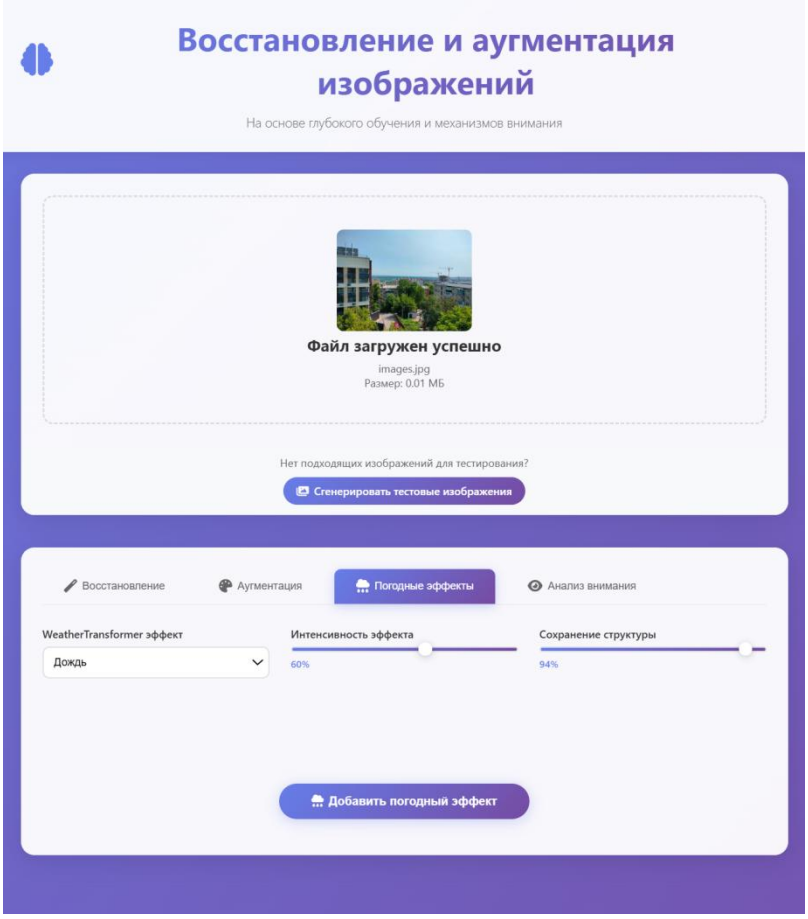


Рисунок 4.11 – Интерфейс программного комплекса

Практическое применение ПК демонстрирует, что эффективное решение задач компьютерного зрения требует понимания двойственности процессов деградации и восстановления. Это открывает новые возможности для создания самообучающихся систем, которые могут улучшать качество работы за счет понимания природы искажений.

Интегрированный подход позволяет создавать более устойчивые системы обработки изображений для реальных приложений, где условия съемки могут значительно варьироваться. Основываясь на принципах, заложенных при

разработке программного комплекса, необходимо детально рассмотреть алгоритмы аугментации данных и способы их применения в задачах компьютерного зрения.

Выводы по главе

1. Представлены и проанализированы различные подходы к синтезу, генерации изображений в целях аугментации изображений, включая алгоритмы зашумления, частичной стилизации с использованием блоков AdaIN и синтеза погодных эффектов. Эксперименты показали, что предложенные методы позволяют существенно повысить точность моделей классификации и сегментации изображений, особенно в задачах с ограниченным числом обучающих данных, что подтверждается результатами на задачах компьютерного зрения с использованием различных обучающих данных.

2. Представлена новая модель глубокой нейронной сети WeatherTransformer, предназначенная для синтеза изображений путем переноса из эталонного и включения в синтезированное атмосферных осадков. Модель основана на объединении сверточных и трансформерных технологий. Модель включает сверточный энкодер и декодер, обеспечивающие извлечение и восстановление признаков на различных масштабах, а также блоки перекрестного внимания, позволяющие точно перенести образец эффекта на исходную сцену. Введена специальная функция потерь, учитывающая одновременно контентное, структурное и сходство восприятия с эталоном, что способствует получению визуально реалистичных погодных осадков без искажения структуры изображения.

Проведенное исследование показало преимущества перекрестного внимания, прогрессивного декодирования и использования шаблонов для управления стилем. Архитектура WeatherTransformer способна синтезировать разнообразные эффекты (туман разной густоты, дождь различных интенсивностей, снег разной структуры) в рамках единой обученной модели, просто меняя входной шаблон.

3. Демонстрируется универсальность предложенного подхода. Несмотря на разную природу осадков, используется одна и та же архитектура нейронной сети, способная без дообучения синтезировать различные их виды. Это открывает возможности для множества приложений: от создания тренировочных данных с разнообразными условиями окружающей среды до инструментов фотомонтажа и киноиндустрии, где требуется добавлять погодные явления уже непосредственно после съёмки.

Полученные результаты показывают, что синтезированные изображения могут использоваться для аугментации данных в задачах восстановления изображений, сегментации и детектирования объектов. Показывается, что использование данного способа аугментации способствует улучшению эффективности работы нейронных сетей по сравнению с использованием классических методов аугментации данных.

4. В результате проведенных исследований разработан программный комплекс восстановления и аугментации изображений, основанный на принципах циклического обучения. Комплекс объединяет модули восстановления, генерации искажений, синтеза погодных эффектов и анализа внимания в единую итеративную систему, где задачи восстановления и генерации, стиллизации выступают как взаимнообратимые процессы. Это позволяет формировать синтетические пары изображений и прогрессивно улучшать качество обеих задач, даже при ограниченном объеме размеченных данных.

Модульная архитектура ПК ВИА обеспечивает гибкость конфигурации и адаптацию под различные условия съемки, включая зашумление, стилизацию и атмосферные искажения. Проведенные эксперименты показали, что использование ПК ВИА позволяет при обработке типовых изображений достичь высоких результатов по метрикам PSNR, SSIM и FID, а также улучшить интерпретируемость моделей благодаря визуализации карт внимания. Программный комплекс продемонстрировал работоспособность и универсальность при решении широкого спектра задач компьютерного зрения и может быть эффективно применен в различных прикладных областях.

Заключение

В ходе выполнения диссертационной работы были получены следующие научные результаты.

1. Проведен анализ современных подходов в задачах восстановления и аугментации изображений. Выявлены ограничения классических методов обработки изображений (линейной и нелинейной фильтрации, частотных преобразований, методов регуляризации), обусловленные их высокой чувствительностью и зависимостью к выбору параметров восстановления. Показано, что классические алгоритмы недостаточно эффективны для восстановления изображений от сложных и нестандартных искажений. Обоснована актуальность применения методов глубокого обучения (сверточных нейронных сетей и трансформеров) благодаря их высокой обобщающей способности и отсутствия необходимости ручной настройки параметров. Установлено, что основным фактором, ограничивающим эффективность нейросетевых алгоритмов восстановления, является нехватка обучающих данных, что делает особенно важным использование продвинутых средств аугментации данных для расширения обучающих выборок и моделирования различных типов шумов и искажений.

2. Проанализированы современные методы аугментации данных, включая как эвристические подходы, так и генеративные модели (GAN, VAE, диффузионные модели). Показано, что генеративные алгоритмы способны достоверно синтезировать различные искажения, близкие к реальным (атмосферные осадки, шумы, артефакты и пр.). Отмечено, что для управляемого переноса стиля и разнообразия синтезированных данных целесообразно использование специализированных нейросетевых архитектур. Выявлена необходимость использования комплексного подхода, объединяющего нейросетевые алгоритмы восстановления изображений с целенаправленной аугментацией данных, как основу для повышения качества изображений в условиях нехватки данных и высокой вариативности искажений.

На основе проведенного анализа разработана общая схема построения и исследования алгоритмов восстановления и аугментации изображений, ставшая методологической основой для диссертации. Данная схема легла в основу архитектуры ПК ВИА, реализующей циклическое обучение, в котором модули восстановления и генерации последовательно улучшают друг друга на основе синтетических и реальных данных.

3. Теоретически обоснованы новые подходы к структурной регуляризации механизма самовнимания с целью повышения устойчивости трансформерных сетей к переобучению. Предложен метод внесения стохастической составляющей в механизм внимания. Показано, что такая регуляризация сглаживает доминирующие весовые коэффициенты внимания и препятствует неконтролируемому росту отдельных весов в процессе обучения. Данный подход позволяет улучшить обобщающую способность модели, повышая ее устойчивость к переобучению.

Предложен и теоретически обоснован способ структурной регуляризации процесса обучения трансформеров, отличающийся использованием обучаемой матрицы масштабных коэффициентов. Теоретически показано, что применение подобной матрицы, которая вносится в схему обработки путем поэлементного перемножения с матрицами скалярных произведений в трансформерных блоках, способствует выходу активационной функции, применяемой в механизме внимания из области насыщения.

4. Разработаны обладающие новизной архитектуры ГНС трансформерного типа для задачи восстановления изображений, отличающиеся использованием канального механизма внимания со сжатием информации, уменьшающим размерность внутреннего представления признаков, а также использованием механизма пространственного внимания с внесением дополнительной стохастической составляющей. Экспериментально показано, что при умеренном сжатии каналов канальной информации качество восстанавливаемых изображений снижается незначительно, тогда как вычислительная сложность блока внимания сокращается пропорционально

коэффициенту сжатия. Для сети с модифицированным пространственным вниманием, опирающейся на базовую архитектуру SwinIR, в ходе экспериментов установлено, что различные варианты структурной регуляризации в рамках общих теоретически обоснованных принципов превосходят исходную модель.

5. Синтезированы и исследованы алгоритмы генерации и стилизации изображений на основе глубоких нейронных сетей для целей аугментации данных. Разработаны практические методики добавления реалистичных искажений: алгоритмы зашумления изображений различными типами шума, частичная стилизация изображений с использованием блока Adaptive Instance Normalization для переноса стилевых характеристик, а также алгоритмы синтеза эффектов плохих погодных условий.

Предложена оригинальная модель WeatherTransformer для добавления атмосферных осадков на изображения реальных сцен. Архитектура WeatherTransformer объединяет сверточный энкодер-декодер для извлечения и восстановления многоуровневых признаков с блоками перекрестного внимания, точно накладывающими шаблон погодного эффекта (дождь, снег, туман) на исходное изображение. При ее обучении введена специальная функция потерь, учитывающая содержательное и структурное сходство с исходным изображением, а также чувствительное сходство (ошибка восприятия) с эталонным образцом эффекта. Показано, что использование механизма перекрестного обеспечивает реалистичный синтез изображений без искажения структуры сцены. Обученная модель WeatherTransformer способна с высокой достоверностью синтезировать различные виды осадков (от легкой дымки или мороси до сильного тумана, дождя и снегопада) в рамках единой архитектуры, что свидетельствует о ее универсальности.

6. Разработан программный комплекс ПК ВИА для восстановления и аугментации изображений, основанный на принципах циклического обучения, в котором модули восстановления и генерации последовательно улучшают друг друга на основе синтетических и реальных данных. Экспериментально подтверждена эффективность использования синтезированных и искаженных

изображений в качестве обучающих данных в задачах восстановления, классификации и сегментации в условиях ограниченных объемов обучающих данных.

Рекомендации по использованию. Разработанные модели и алгоритмы улучшения качества изображений могут быть внедрены в различные системы компьютерного зрения и цифровой обработки изображений. Наиболее перспективными областями применения являются автоматизированные системы, требующие восстановления или повышения качества визуальных данных в присутствии шумов и помех: системы видеонаблюдения и безопасности (для очистки и детализации кадров); системы дистанционного аэрокосмического мониторинга (улучшение спутниковых снимков в сложных погодных условиях), медицинская визуализация (подавление артефактов и улучшение резкости снимков), мобильная фотография и постобработка изображений.

Предложенные алгоритмы аугментации данных могут использоваться для расширения тренировочных наборов при обучении нейронных сетей различной архитектуры, что особенно важно в задачах, где сбор реальных зашумленных данных затруднен или объем данных ограничен. Например, генерация синтетических дождя или тумана на изображениях позволит подготовить систему компьютерного зрения к работе в реальных погодных условиях без необходимости ждать наступления таких условий для фотосъемки, а добавление контролируемого шума – повысить устойчивость модели к соответствующим помехам в эксплуатации.

Перспективы дальнейшего развития. Результаты работы открывают ряд направлений для будущих исследований. Одним из таких направлений является дальнейшее совершенствование гибридных нейросетевых алгоритмов путем интеграции предложенных механизмов внимания с другими современными архитектурными решениями и методами аугментации. Перспективным также является реализация подхода к совместному обучению нескольких нейронных сетей, задействованных на разных этапах восстановления и аугментации, на

единых данных, что позволит более точно и согласованно настраивать их параметры, чтобы добиться лучшей производительности и эффективности.

Кроме того, важным направлением развития исследования является применение разработанных моделей и алгоритмов к обработке видеоданных (потоковое видео). Требуется более детальное рассмотрение вопросов регуляризации внимания и аугментации в динамических сценах, например, синтез и обработка помех с учетом их временной корреляции на соседних кадрах видеопоследовательности. Такой подход позволил бы эффективно восстанавливать видеоизображения даже при длительных или зависимых во времени искажениях.

Список использованных источников

1. Zhang J. Quantile analysis of image sensor noise distribution / J. Zhang, K. Hirakawa, X. Jin // ICASSP. – 2015. – DOI: 10.1109/ICASSP.2015.7178240.

2. Сизиков В.С. Устойчивые методы обработки результатов измерений / В.С. Сизиков. – Санкт-Петербург.: СпецЛит, 1999. – 240 с.

3. Heckel R., Soltanolkotabi M. Denoising and regularization via exploiting the structural bias of convolutional generators //arXiv preprint arXiv:1910.14634. – 2019.

4. Бережнов Н.И. Исследование обобщающей способности методов глубокого обучения для улучшения качества изображений / Н.И. Бережнов, А.А. Сирота // XXIII Международная конференция «Информатика: проблемы, методология, технологии». – Воронеж: ИПЦ ВГУ: – 2023. – С. 510-518.

5. Wang Z. Defect simulation in SEM images using generative adversarial networks / Z. Wang, Y. Liangjiang, P. Lingling // SPIE Advanced Lithography. – 2021. – DOI: 10.1117/12.2581881.

6. Zuo W. Texture Enhanced Image Denoising via Gradient Histogram Preservation / W. Zuo, L. Zhang, C. Song, D. Zhang // CVPR. – 2013. – DOI: 10.1109/CVPR.2013.159.

7. Бондина Н.Н. Адаптивные алгоритмы фильтрации и изменения контраста изображения / Н.Н. Бондина, Р.Ю. Мураров // Вестник НТУ. – 2014. – №35. – 8 с.

8. Гонсалес Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М.: Техносфера, 2012. – 1104 с.

9. Тихонов В.И. Статический анализ и синтез радиотехнических устройств и систем связи / В.И. Тихонов, В.Н. Харисов. – М.: Радио и связь, 2004. – 608 с.

10. Клочко В.К. Методы восстановления изображений и оценивания аппаратной функции по прореженной матрице наблюдений / В.К. Клочко, В.П. Кузнецов // Автометрия. – 2016. – Т. 52. – №6. – С. 12–20. – DOI: 10.15372/AUT20160602.

11. Milukova O. Image Restoration Spectral Techniques/ O. Milukova, V. Kober, I.A. Ovseevich // PRIP. – 2009. – 4 с.

12. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.

13. Ваняшкин Ю.Ю. Применение автокодировщиков для устранения шумов с изображений / Ю.Ю Ваняшкин, Д.А. Макаров // Научно-образовательный журнал для студентов и преподавателей «StudNet». – 2020. – №10. – 8 с.

14. Szegedy C., Liu W., Jia Y., et al. Going deeper with convolutions [Электронный ресурс] // arXiv preprint arXiv:1409.4842. – 2014. – Режим доступа: https://arxiv.org/abs/1409.4842.

15. Chollet F. Xception: Deep learning with depthwise separable convolutions [Электронный ресурс] // arXiv preprint arXiv:1610.02357. – 2016. – Режим доступа: https://arxiv.org/abs/1610.02357.

16. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition [Электронный ресурс] // arXiv preprint arXiv:1512.03385. – 2015. – Режим доступа: https://arxiv.org/abs/1512.03385.

17. Xie S., Girshick R., Dollár P., Tu Z., He K. Aggregated residual transformations for deep neural networks [Электронный ресурс] // arXiv preprint arXiv:1611.05431. – 2016. – Режим доступа: https://arxiv.org/abs/1611.05431.

18. Huang G., Liu Z., Van Der Maaten L., Weinberger K.Q. Densely connected convolutional networks [Электронный ресурс] // arXiv preprint arXiv:1608.06993. – 2016. – Режим доступа: https://arxiv.org/abs/1608.06993.

19. Tan M., Le Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks [Электронный ресурс] // arXiv preprint arXiv:1905.11946. – 2019. – Режим доступа: https://arxiv.org/abs/1905.11946.

20. Liu Z., Mao H., Wu C.-Y., et al. A ConvNet for the 2020s [Электронный ресурс] // arXiv preprint arXiv:2201.03545. – 2022. – Режим доступа: https://arxiv.org/abs/2201.03545.

21. Xiao-Jiao M. Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections / M. Xiao-Jiao, S. Chunhua, Y. Yubin // NIPS. – 2016.

22. Deya B. SEM image denoising with Unsupervised Machine Learning for better defect inspection and metrology / B. Deya, S. Haldera, K. Khalil // SPIE Advanced Lithography. – 2021. – DOI: 10.1117/12.2584803.

23. Liu J., Lin Y., Hu J., et al. IFSR-Net: Image restoration using implicit frequency selection and recovery // Machine Vision and Applications. – 2025. – Vol. 36, №1. – DOI: 10.1080/09540091.2025.2465448.

24. Gupta S., Sharma P., Agarwal S., et al. CV-CAN and CV-DDAN: Complex-valued attention networks for image denoising and restoration // Frontiers in Artificial Intelligence. – 2024. – Vol. 7. – Article 1353873. – DOI: 10.3389/frai.2024.1353873.

25. Liu Z., Zhou Y., Han X., et al. VmambaIR: Visual state space model for image restoration // arXiv preprint. – 2024. – arXiv:2403.11423. – Режим доступа: https://arxiv.org/abs/2403.11423.

26. Chen C., Zhang J., Li X., et al. KBNet: Kernel-based attention network for image restoration // arXiv preprint. – 2023. – arXiv:2303.02881. – Режим доступа: https://arxiv.org/abs/2303.02881.

27. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale // arXiv preprint. – 2020. – № arXiv:2010.11929.

28. Cordonnier J., Loukas A., Jaggi M. On the Relationship between Self-Attention and Convolutional Layers // arXiv preprint. – 2019. – № arXiv:1911.03584. – DOI: 10.48550/arXiv.1911.03584.

29. Zhao H., Jia J., Koltun V. Exploring Self-Attention for Image Recognition // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 10076–10085.

30. Liang J., Cao J., Sun G., Zhang K., Van Gool L., Timofte R. SwinIR: Image Restoration Using Swin Transformer // Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021. P. 1833–1844. DOI: 10.1109/ICCVW54120.2021.00058.

31. Zhang J., Qin Q., Ye Q., Ruan T. ST-UNet: Swin Transformer Boosted U-Net with Cross-Layer Feature Enhancement for Medical Image Segmentation // Computers in Biology and Medicine. – 2023. – Vol. 153. – DOI: 10.1016/j.compbiomed.2022.106516.

32. Illarionova S., Shadrin D., Shukhratov I., Evteeva K., Popandopulo G., Sotiriadi G., Burnaev E. Benchmark for Building Segmentation on Up-Scaled Sentinel-2 Imagery // Remote Sensing. – 2023. – Vol. 15, № 9. – Article ID: 2347. – DOI: 10.3390/rs15092347.

33. Xie E., Wang W., Yu Z., Anandkumar A., Alvarez J.M., Luo P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers // arXiv preprint. – 2021. – № arXiv:2105.15203. – DOI: 10.48550/arXiv.2105.15203.

34. Fan C.-M., Lin T.-J., Lin K.-H. SUNet: Swin Transformer UNet for Image Denoising // Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS). – 2022. – DOI: 10.1109/ISCAS48785.2022.9937486.

35. Wang C., Pan J., Wu X. Structural Prior Guided Generative Adversarial Transformers for Low-Light Image Enhancement // arXiv preprint. – 2022. – № arXiv:2207.07828. – DOI: 10.48550/arXiv.2207.07828.

36. Zamir S.W., Arora A., Khan S., Hayat M., Khan F.S., Yang M. Restormer: Efficient Transformer for High-Resolution Image Restoration // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2022. – P. 5728–5739.

37. Valanarasu J.M., Yasarla R., Patel V.M. TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2022. – P. 2353–2363.

38. Jing L., Tian Y. Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2021. – Vol. 43, № 11. – P. 4037–4058.

39. Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., Xiong H., He Q. A Comprehensive Survey on Transfer Learning // Proceedings of the IEEE. – 2021. – Vol. 109, № 1. – P. 43–76. – DOI: 10.1109/JPROC.2020.3004555.

40. Бережнов Н.И. Универсальный алгоритм повышения качества изображений с использованием глубоких нейронных сетей / Н.И. Бережнов, А.А. Сирота // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2022. – № 2. – С. 81–92. – DOI: 10.17308/sait/1995-5499/2022/2/81-92.

41. Ali A., Benjdira B., Bazi Y., Koubaa A. Vision Transformers in Image Restoration: A Survey // Sensors. – 2023. – Vol. 23, № 5. – Article ID: 2385. – DOI: 10.3390/s23052385.

42. Xie Q. Unsupervised Data Augmentation for Consistency Training / Q. Xie, Z. Dai1, E. Hovy, M. Luong, Q. V. Le // arXiv. – 2020. Режим доступа: https://arxiv.org/abs/1904.12848.

43. Klinger R. Classical Probabilistic Models and Conditional Random Fields / R. Klinger, A. Tomanek // Algorithm Engineering Report TR07-2-013. of Computer Science. – Dortmund University of Technology, 2007.

44. Krizhevsky A. ImageNet Classification with Deep Convolutional Neural Networks / A. Krizhevsky, I. Sutskever, E. Geoffrey // Proceedings of the 25th International Conference on Neural Information Processing Systems. – 2012 – Vol. 1 – P. 1097-1105.

45. Gayer A.V. Effective real-time augmentation of training dataset for the neural networks learning / A.V. Gayer, Y.S. Chernyshova, A.V. Sheshkus // International Conference on Machine Vision. – 2019.

46. Xu M. et al. A comprehensive survey of image augmentation techniques for deep learning // Pattern Recognition. – 2023. – Т. 137. – С. 109347.

47. Shorten C., Khoshgoftaar T.M. A survey on Image Data Augmentation for Deep Learning // Journal of Big Data. – 2019. – Vol. 6, №1. – P. 1–48. – DOI: 10.1186/s40537-019-0197-0.

48. Cubuk E.D., Zoph B., Mane D., Vasudevan V., Le Q.V. AutoAugment: Learning Augmentation Policies from Data // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 113–123. – Режим доступа: https://arxiv.org/abs/1805.09501.

49. Cubuk E.D., Zoph B., Shlens J., Le Q.V. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space // arXiv preprint. – 2020. – arXiv:1909.13719. – Режим доступа: https://arxiv.org/abs/1909.13719.

50. Goodfellow I. NIPS 2016 Tutorial: Generative Adversarial Networks / I. J. Goodfellow // arXiv. – 2017. Режим доступа: http://arxiv:org/abs/1701:00160.

51. Емельянов С.О. Методы аугментации обучающих выборок в задачах классификации изображений / С.О. Емельянов, А.А. Иванова, Е.А. Швец, Д.П. Николаев // Сенсорные системы. – 2018. – Т. 32. – № 3.

52. Doersch C. Tutorial on Variational Autoencoders / C. Doersch – 2016.

53. Oord A. Pixel Recurrent Neural Networks. / A. Oord, N. Kalchbrenner, K. Kavukcuoglu. – 2016.

54. Niu S. et al. Defect image sample generation with GAN for improving defect recognition //IEEE Transactions on Automation Science and Engineering. – 2020. – Т. 17. – №. 3. – С. 1611-1622.

55. Donahue J., Simonyan K. Large scale adversarial representation learning //Advances in neural information processing systems. – 2019. – Т. 32.

56. Shuanlong N. Defect Image Sample Generation with GAN for Improving Defect Recognition / N. Shuanlong, L. Bin, W. Xinggang, L. Hui // IEEE Transactions on Automation Science and Engineering. – 2020. – P. 1-12.

57. Dhariwal P., Nichol A. Diffusion models beat gans on image synthesis //Advances in neural information processing systems. – 2021. – Т. 34. – С. 8780-8794.

58. Lucic M. Are GANs Created Equal? A Large-Scale Study / M. Lucic, K. Kurach, M. Michalski, S. Gelly, O. Bousquet // arXiv: 1711.10337 – 2017.

59. Dhariwal P. Diffusion Models Beat GANs on Image Synthesis / P. Dhariwal, A. Nichol // arXiv: 2105.05233 – 2021.

60. Isola P. Image-to-Image Translation with Conditional Adversarial Networks / P. Isola, J. Zhu, T. Zhou, A. A. Efros // arXiv preprint. – 2016. – arXiv:1611.07004.

61. Zhu J.-Y. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks / J.-Y. Zhu, T. Park, P. Isola, A. A. Efros // arXiv preprint. – 2017. – arXiv:1703.10593.

62. Huang X. Multimodal Unsupervised Image-to-Image Translation / X. Huang, M.-Y. Liu, S. Belongie, J. Kautz // arXiv preprint. – 2018. – arXiv:1804.04732. – DOI: 10.48550/arXiv.1804.04732.

63. Lee H.-Y. DRIT++: Diverse Image-to-Image Translation via Disentangled Representations / H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, M.-H. Yang // arXiv preprint. – 2020. – arXiv:1905.01270. – DOI: 10.48550/arXiv.1905.01270.

64. Zhou K. High-resolution Rainy Image Synthesis: Learning from Rendering / K. Zhou, S. Zhao, H. Deng, L. Zhang // arXiv preprint. – 2025. – № arXiv:2502.16421.

65. Wei M., Shen Y., Wang Y., Xie H., Qin J., Wang F. L. RainDiffusion: When Unsupervised Learning Meets Diffusion Models for Real-world Image Deraining // Nanjing University of Aeronautics and Astronautics; Lingnan University; The Hong Kong Polytechnic University; Hong Kong Metropolitan University. – 2024.

66. Parmar P., Kundurthy S., Lee Y. One-Step Image Translation with Text-to-Image Models (CycleGAN-Turbo) // arXiv preprint. – 2024. – № arXiv:2403.12036.

67. Zhang L. Adding Conditional Control to Text-to-Image Diffusion Models / L. Zhang, A. Rao, M. Agrawala // arXiv preprint. – 2023. – arXiv:2302.05543. – DOI: 10.48550/arXiv.2302.05543.

68. Greenberg A., Elidan G., Shocher A. Seed-to-Seed: Image Translation in Diffusion Seed Space // arXiv preprint. – 2024.

69. Qian C., Lin Y., Zhang X., et al. WeatherDG: LLM-assisted Diffusion Model for Procedural Weather Generation // arXiv preprint. – 2024.

70. Pang L., Liu Y., Yang Y., Zhang Y. TRG-Net: An Interpretable and Controllable Rain Generator // arXiv preprint. – 2024. – № arXiv:2403.09993.

71. Wang C., Li Y., Chen J., et al. Mask-DerainGAN: Learning to remove rain streaks by learning to generate rainy images / Wang C., Li Y., Chen J., et al // Pattern Recognition. – 2024. – Vol. 156.

72. Ali A. Xcit: Cross-covariance image transformers / A. Ali et al. // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 20014–20027.

73. Chen B. Psvit: Better vision transformer via token pooling and attention sharing / B. Chen [et al.] // arXiv preprint arXiv:2108.03428. – 2021.

74. Yuan L. Volo: Vision outlooker for visual recognition / L. Yuan et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2022. – Vol. 45, № 5. – P. 6575–6586.

75. Zhang D. Swinfir: Revisiting the swinir with fast Fourier convolution and improved training for image super-resolution / D. Zhang // arXiv preprint arXiv:2208.11247. – 2022.

76. Zhao H. Comprehensive and delicate: An efficient transformer for image restoration / H. Zhao et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2023. – P. 14122–14132.

77. Xia Z. DAT++: Spatially Dynamic Vision Transformer with Deformable Attention / Z. Xia et al. // arXiv preprint arXiv:2309.01430. – 2023.

78. Ren B. Key-Graph Transformer for Image Restoration / B. Ren et al. // arXiv preprint arXiv:2402.02634. – 2024.

79. Wang C. How Powerful Potential of Attention on Image Restoration? / C. Wang et al. // arXiv preprint arXiv:2403.10336. – 2024.

80. Chen X. Activating more pixels in image super-resolution transformer / X. Chen et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2023. – P. 22367–22377.

81. Gao H. Prompt-based Ingredient-Oriented All-in-One Image Restoration / H. Gao et al. // IEEE Transactions on Circuits and Systems for Video Technology. – 2024.

82. Мэрфи, К. П. Вероятностное машинное обучение. Введение. MIT Press, 2022.

83. Yeh C., Chen Y., Wu A., Chen C., Viégas F., Wattenberg M. AttentionViz: A Global View of Transformer Attention [Электронный ресурс] // arXiv preprint arXiv:2305.03210. – 2023. – Режим доступа: https://arxiv.org/abs/2305.03210.

84. Li Y., Wang J., Dai X., Wang L., Yeh C.C.M., Zheng Y., Ma K.L. How Does Attention Work in Vision Transformers? A Visual Analytics Attempt // IEEE Transactions on Visualization and Computer Graphics. – 2023. – DOI: 10.1109/TVCG.2023.3242584.

85. Lu Y., Lin Y., Wu H., Luo Y., Zheng X., Wang L. All one needs to know about priors for deep image restoration and enhancement: A survey [Электронный ресурс] // arXiv preprint arXiv:2206.02070. – 2022. – Режим доступа: https://arxiv.org/abs/2206.02070.

86. Бережнов Н.И. Влияние априорной информации на механизм внимания взадаче улучшения качества изображений в моделях-трансформерах / Н.И. Бережнов, А.А. Сирота // XXIV Международная конференция «Информатика: проблемы, методология, технологии». – Воронеж: ИПЦ ВГУ: 2024. – С. 602-609.

87. Jetley S., Lord N.A., Lee N., Torr P.H.S. Learn to pay attention [Электронный ресурс] // arXiv preprint arXiv:1804.02391. – 2018. – Режим доступа: https://arxiv.org/abs/1804.02391.

88. Huynh-Thu Q., Ghanbari M. Scope of validity of PSNR in image/video quality assessment // Electronics Letters. – 2008. – Vol. 44, № 13. – P. 800–801. – DOI: 10.1049/el:20080522.

89. Berezhnov N.I. Understanding the attention mechanism in neural network transformer models in image restoration tasks / N.I. Berezhnov, A.A. Sirota // 5nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA). – Lipetsk: 2023. – P. 207-211.

90. Zhang H., Qu D., Shao K., Yang X. Dropdim: A regularization method for transformer networks // IEEE Signal Processing Letters. 2022. Vol. 29. P. 474-478.

91. Zhou W., Ge T., Xu K., Wei F., Zhou M. Scheduled drophead: A regularization method for transformer models // arXiv preprint arXiv:2004.13342. 2020.

92. Zehui L., Liu P., Huang L., Chen J., Qiu X., & Huang, X. (2019). Dropattention: A regularization method for fully-connected self-attention networks. arXiv preprint arXiv:1907.11065.

93. Li B., Hu Y., Nie X., Han C., Jiang X., Guo T., Liu L. Dropkey for vision transformer // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. P. 22700-22709.

94. Розанов Ю.А. Случайные поля и стохастические уравнения с частными производными. – М.: Наука, 1977.

95. Бережнов Н.И. Модификации механизмов внимания в моделях-трансформерах в задаче восстановления изображений / Н.И. Бережнов // XXV Международная конференция «Информатика: проблемы, методология, технологии». – Воронеж: ИПЦ ВГУ: 2025.

96. Бережнов Н.И. Регуляризация механизма самовнимания в блоках трансформеров и ее применение в задачах классификации и восстановления изображений / Н.И. Бережнов, А.А. Сирота // Искусственный интеллект и принятие решений. – М.: – 2025. – №2. – С. 114-129.

97. Abdelhamed A. A high-quality denoising dataset for smartphone cameras / A. Abdelhamed, S. Lin, M.S. Brown // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – P. 1692–1700.

98. Buslaev A. Albumentations: Fast and Flexible Image Augmentations / A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov // Information. 2020. – Vol. 11. – P. 125. DOI: https://doi.org/10.3390/info11020125.

99. Бережнов Н.И. Совершенствование механизмов внимания для архитектуры трансформер в задачах повышения качества изображений / Н.И. Бережнов, А.А. Сирота // Компьютерная оптика. – Самара: – 2024. – Т. 48. – №. 5. – С. 726-733.

100. Zamir S.W. Learning enriched features for real image restoration and enhancement / S.W. Zamir // Computer Vision–ECCV 16th European Conference. – 2020. – Vol. 16. – P. 492-511.

101. Алгазинов Э. К. Анализ и компьютерное моделирование информационных процессов и систем / Э. К. Алгазинов, А. А. Сирота. – М.: Диалог-МИФИ, 2009. – 416 с.

102. Сирота А.А. Анализ потенциальных и реальных характеристик оценивания случайных полей (изображений) в условиях аддитивных и импульсных помех / А.А. Сирота, П.В. Калинин // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2011. – №1 – С. 41-50.

103. Shorten C. A survey on Image Data Augmentation for Deep Learning / C. Shorten, T. Khoshgoftaar. // Journal of Big Data. – 2019.

104. Neural Style Transfer: Applications in Data Augmentation [Электронный ресурс]. Режим доступа: https://towardsdatascience.com/neural-style-transfer-applications-data-augmentation-43d1dc1aeecc.

105. Бережнов Н.И. Модели глубокого обучения для синтеза изображений с включением атмосферных осадков с целью решения задач компьютерного зрения в различных погодных условиях / Н.И. Бережнов, А.А. Сирота // Вестн. Воронежского гос. ун-та, Сер. Системный анализ и информационные технологии. – Воронеж: ИПЦ ВГУ: – 2025. – № 2. – С. 89-104.

106. Smith L. N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates / L. N. Smith, N. Topin // arXiv preprint. – 2017. – arXiv:1708.07120. Режим доступа: https://arxiv.org/abs/1708.07120.

107. Ronneberger O. U-Net: Convolutional Networks for Biomedical Image Segmentation / O. Ronneberger, P. Fischer, T. Brox. – 2015.

108. Свидетельство о государственной регистрации программы для ЭВМ 2025682168 Российская Федерация. Программный комплекс для автоматизированного восстановления и аугментации графических данных / Н.И. Бережнов; заявитель и правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный университет». − № 2025682168; заявление 07.08.2025; опубл. 21.08.2025.

Приложение А. Акты о внедрении

УТВЕРЖДАЮ
Декан факультета компьютерных наук
_____Крыловецкий А.А.
«04» сентября 2025г.

АКТ
о внедрении (использовании) результатов
кандидатской диссертационной работы
Бережнова Никиты Игоревич

Комиссия в составе:
председатель к.ф.-м.н., доц. А.А. Крыловецкий
члены комиссии: к.т.н., доц. Е.Ю. Митрофанова, к.т.н. Иванков А.Ю.

составили настоящий акт о том, что результаты диссертационной работы Бережнова Никиты Игоревича «Совершенствование механизмов внимания в глубоких нейронных сетях – трансформерах в задачах восстановления и аугментации изображений», представленной на соискание ученой степени кандидата технических наук, использованы в учебном процессе Факультета компьютерных наук на кафедре технологий обработки и защиты информации при проведении лекционных и практических занятий по следующим дисциплинам: «Разработка приложений для систем машинного обучения», «Проектный семинар «Машинное обучение», а также и подготовки выпускных квалификационных работ.

При подготовке и проведении занятий нашли отражения следующие результаты диссертационной работы Бережнова Никиты Игоревича:

- методы реализации и модификации механизмов внимания в трансформерных модулях глубоких нейронных сетей при обработке изображений;
- способы снижения вычислительной сложности модулей внимания в трансформерах за счет использования предложенного алгоритма канального сжатия;
- предложенный подход к аугментации изображений с целью учета факторов, негативных для восприятия сцен и, прежде всего, атмосферных осадков (дождь, снег, туман).

Председатель комиссии:
Декан факультета компьютерных наук,
к.ф-м.н, доцент — Крыловецкий А.А.

Члены комиссии:
Доцент кафедры технологий обработки
и защиты информации, к.т.н., доцент — Митрофанова Е.Ю.

Доцент кафедры технологий обработки
и защиты информации, к.т.н. — Иванков А.Ю.

УТВЕРЖДАЮ
И.О. проректора по науке,
инновациям и цифровизации
ФГБОУ ВО «Воронежский
государственный университет»
д.ф.-м.н., доцент
Костин Д.В.
«09» 2025г.

Справка об использовании

результатов диссертационной работы Бережнова Н.И. на тему «Совершенствование механизмов внимания в глубоких нейронных сетях – трансформерах в задачах восстановления и аугментации изображений»

В период обучения в аспирантуре и подготовки диссертационной работы Бережнов Никита Игоревич принимал участие в выполнении следующих НИОКР, проводимых ФГБОУ ВО «ВГУ» в период 2021-2025г.г.: СЧ НИР НИЧ № 21009, гос. контракт № 70 /2021, СЧ НИР НИЧ № 23019, гос. контракт № 47/2023.

Результаты его диссертационного исследования использованы в указанных НИОКР в части:

- разработки и исследования нейросетевых алгоритмов повышения качества изображений специального вида;
- разработки и исследования алгоритмов аугментации и стилизации изображений для целей расширения объемов обучающих данных нейросетевых алгоритмов обработки информации.

Справка выдана для представления в диссертационный совет по месту защиты.

Заместитель научного руководителя, ответственный исполнитель СЧ НИР
Кандидат технических наук, доцент

«04» сентября 2025 г. М.А. Дрюченко

Приложение Б. Свидетельство о государственной регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025682168

Программный комплекс для автоматизированного восстановления и аугментации графических данных

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный университет» (RU)***

Автор(ы): ***Бережнов Никита Игоревич (RU)***

Заявка № **2025680273**
Дата поступления **07 августа 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***21 августа 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов