DOI: 10.15514/ISPRAS-2026-38(6)-36

CC BY

# Исследование применения диффузионного редактирования для аугментации в задачах классификации изображений

[1,2] Д.А. Матюшин, ORCID: 0009-0003-6737-0129 <matiushin.da@ispras.ru>
[1] О.И. Быков, ORCID: 0009-0007-7151-0533 <olbkv@ispras.ru>
[1] И.И. Бубенков, ORCID: 0009-0001-5134-5098 <i.bubenkov@ispras.ru>

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

[2] Московский физико-технический институт,
Россия, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9.

**Аннотация.** Современные задачи классификации изображений часто ограничены дисбалансом классов, малым объёмом данных и недостаточной репрезентативностью выборки, что существенно снижает обобщающую способность моделей. Традиционные методы аугментации (геометрические трансформации, изменение цвета) вносят лишь локальные, семантически ограниченные изменения. В работе представлено систематическое исследование применения редактирования изображений на основе визуально-языковых моделей (VLM) и диффузионных архитектур как эффективного подхода к семантической аугментации. Проведен сравнительный анализ двух современных конвейеров VLM-редактирования изображений (Instruct-Pix2Pix и Qwen-Image-Edit) с классическими и современными диффузионными методами аугментации на наборах данных CUB-200-2011 (с искусственным дисбалансом) и 13-классовом подмножестве iWildCam (с естественным дисбалансом). Разработаны пять стратегий селективного отбора изображений различной адаптивности к структуре дисбаланса, а также оценено влияние типов текстовых инструкций и ключевых гиперпараметров системы. Полученные результаты подтверждают высокую эффективность семантически насыщенного VLM-редактирования для повышения устойчивости классификаторов в условиях выраженного дисбаланса классов.

**Ключевые слова:** аугментация данных; диффузионные модели; визуально-языковые модели; редактирование изображений; дисбаланс классов; классификация изображений.

**Для цитирования:** Матюшин Д.А., Быков О.И., Бубенков И.И. Исследование применения диффузионного редактирования для аугментации в задачах классификации изображений. Труды ИСП РАН, 2026, том 38, вып. 6, часть 2, стр. 249–274. DOI: 10.15514/ISPRAS–2026–38(6)–36.

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





# Investigation of diffusion-based editing applied to data augmentation in image classification tasks

[1,2] D.A. Matiushin, ORCID: 0009-0003-6737-0129 <matiushin.da@ispras.ru>
[1] O.I. Bykov, ORCID: 0009-0007-7151-0533 <olbkv@ispras.ru>
[1] I.I. Bubenkov, ORCID: 0009-0001-5134-5098 <i.bubenkov@ispras.ru>

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

[2] Moscow Institute of Physics and Technology,
Russia, 141701, Moscow Region, Dolgoprudny, Institutsky Lane, 9.

**Abstract.** Modern image classification tasks are often constrained by class imbalance, limited data volume, and insufficient sample representativeness, all of which significantly reduce the generalization ability of models. Traditional augmentation methods (such as geometric transformations and color modifications) introduce only local, semantically limited variations. This work presents a systematic study of image editing based on vision-language models (VLMs) and diffusion architectures as an effective approach to semantic augmentation. A comparative analysis is conducted between two contemporary VLM-based image editing pipelines (Instruct-Pix2Pix and Qwen-Image-Edit) and both classical and modern diffusion-based augmentation methods on the CUB-200-2011 dataset (with artificially induced imbalance) and a 13-class subset of iWildCam (with natural imbalance). Five strategies for selective image sampling with varying degrees of adaptability to imbalance structure are developed, and the impact of different types of textual instructions as well as key system hyperparameters is evaluated. The results demonstrate the high effectiveness of semantically rich VLM-based editing in improving the robustness of classifiers under conditions of pronounced class imbalance.

**Keywords:** data augmentation; diffusion models; visual-language models; image editing; class imbalance; image classification.

**For citation:** Matiushin D.A., Bykov O.I., Bubenkov I.I. Investigation of diffusion-based editing applied to data augmentation in image classification tasks. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 6, part 2, pp. 249-274 (in Russian). DOI: 10.15514/ISPRAS-2026-38(6)-36.

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

## *1. Введение*

Задачи классификации изображений с использованием глубоких нейронных сетей остаются одной из ключевых областей компьютерного зрения и находят применение в автономном вождении, медицинской диагностике [1], экологическом мониторинге, биологической таксономии, сельском хозяйстве, системах безопасности и видеонаблюдения, промышленном контроле качества, робототехнике, анализе культурного наследия, астрономии и во многих других областях [2-3]. Качество и устойчивость моделей для таких задач в значительной степени определяются объёмом, сбалансированностью и репрезентативностью обучающего набора данных. В реальных условиях исследователи часто сталкиваются с выраженным дисбалансом классов [4], малым размером выборки [5-6] и недостаточным разнообразием примеров внутри классов [7]. Эти факторы приводят к переобучению [7], снижению обобщающей способности и нестабильности предсказаний на тестовых данных, особенно на редких или слабо представленных классах.

Традиционные методы аугментации данных [8] позволяют искусственно увеличить объём выборки, но вносят преимущественно локальные геометрические или фотометрические преобразования [9-10]. Они зачастую сохраняют семантическую близость к исходным изображениям и, как правило, слабо компенсируют дефицит внутриклассового разнообразия

и дисбаланс классов. В результате распределение синтетических данных остаётся близким к исходному, что ограничивает способность модели к обобщению [10-12].

В последние годы значительный прогресс в генерации изображений, в том числе при помощи диффузионных моделей [13-15], открыл новые возможности для контролируемого синтеза данных с выраженными семантическими изменениями [11, 16]. Диффузионные архитектуры, дополненные возможностями визуально-языковых моделей (Visual Language Model, VLM) [17-19], позволяют редактировать изображения по текстовому описанию (prompt) [18, 20-21], сохраняя при этом принадлежность объекта к исходному классу [22]. Такой подход обеспечивает существенно более высокое семантическое разнообразие по сравнению с классическими трансформациями и уже продемонстрировал свою эффективность в ряде работ по автоматической аугментации [22-25].

Настоящая работа посвящена систематическому исследованию применения VLM-редактирования для аугментации наборов данных в условиях дисбаланса классов, а также сравнительному анализу предложенного VLM-подхода с другими известными стратегиями аугментации на двух реальных наборах: CUB-200-2011 [26] (с искусственно индуцированным дисбалансом) и 13-классовом подмножестве набора данных iWildCam [27] (с естественным дисбалансом).

Основные вклады данной работы:

1. Получены количественные результаты сравнительного анализа применения VLM для редактирования изображений (Instruct-Pix2Pix [18] и Qwen-Image-Edit [19]) с другими методами аугментации на двух реальных наборах данных компьютерного зрения с искусственно созданным и естественным дисбалансом. Продемонстрировано превосходство предложенного VLM-подхода над стандартными аугментациями за счет более выраженных семантических изменений в данных.
2. Проведена эмпирическая оценка пяти стратегий селективного отбора изображений для аугментации, обладающих разной степенью адаптивности к структуре дисбаланса набора данных. Получены количественные характеристики влияния таких стратегий на итоговое качество классификации на примере набора CUB, выявлено, что стратегия играет ключевую роль для успеха аугментации.
3. Исследовано влияние типа редактирующих текстовых инструкций (изменение фона, положения объекта, комбинированные изменения) и других гиперпараметров на итоговое качество классификатора; выявлены наиболее оптимальные условия получения прироста качества на исследуемых наборах данных.
4. Дополнительно исследовано влияние автоматической фильтрации синтетической добавки к данным перед обучением классификатора. Показано, что фильтрация может давать дополнительный прирост качества, однако её эффективность зависит от набора данных, модели редактирования и используемого критерия отбора синтетических изображений.

## *2. Обзор литературы*

### 2.1 Диффузионные модели: принцип работы и особенности архитектуры

Прорыв в развитии генеративных моделей на основе диффузионных процессов произошел в 2020 году с публикацией работы [13]. Авторы формализовали прямой диффузионный процесс как марковскую цепь, последовательно добавляющую на изображение гауссовский шум. При большом числе шагов конечное распределение приближается к изотропному гауссовскому шуму. Обратный процесс параметризуется нейронной сетью, которая на





каждом шаге предсказывает наложенный на изображение шум. Модели DDPM впервые продемонстрировали качество генерации, сопоставимое или превосходящее GAN [28], при значительно большей стабильности обучения.

Дальнейшее развитие включало ускорение генерации (DDIM) [14], совершенствование «расписания зашумления» и применение бесклассификаторного управления генерацией (classifier-free guidance) [29], что позволило эффективно управлять генерацией по текстовым условиям. В качестве основной архитектуры в таких моделях часто используется U-Net [30] с пропускающими соединениями (skip-соединениями) и блоками внимания [31].

Существенным ограничением пиксельных диффузионных моделей оставалась высокая вычислительная стоимость при работе с изображениями высокого разрешения. Это ограничение было преодолено в работе [15], где диффузионный процесс был перенесён в латентное пространство предварительно обученного вариационного автокодировщика (VAE). Такой подход существенно снизил требования к памяти и времени вычислений и лёг в основу модели Stable Diffusion, которая добавила к диффузионным моделям текстовую модальность при помощи текстового энкодера CLIP [32] и стала одной из наиболее широко используемых открытых моделей для генерации изображений по тексту. В 2025 году компания Alibaba представила семейство моделей Qwen-Image [19], в частности специализированную модель Qwen-Image-Edit на базе мультимодального диффузионного трансформера (Multimodal Diffusion Transformer, MMDiT) [33], в основе которого лежит принцип двойного кодирования: один канал, управляемый мультимодальной моделью Qwen-VL [34], отвечает за понимание и сохранение семантики изображения, а второй, с помощью вариационного автокодировщика (Variational autoencoder, VAE) [35], фокусируется на точной реконструкции визуальных деталей и текстур. Это позволяет выполнять точное редактирование существующих изображений при сохранении реалистичности и семантической согласованности, делая Qwen-Image-Edit удобным инструментом для задач аугментации данных.

### 2.2 Методы аугментации изображений при помощи диффузионных моделей

В последние годы диффузионные генеративные модели стали важным инструментом расширения обучающих выборок при дефиците размеченных данных [11, 23, 36-37], позволяя синтезировать семантически согласованные и разнообразные изображения. В систематическом исследовании UniDiffDA [24] диффузионная аугментация формализована как последовательный процесс из трех ключевых компонентов: (1) дообучение диффузионной модели (опционально), (2) генерация синтетических выборок и (3) объединение сгенерированных данных с исходной выборкой или замещение реальных образцов синтетическими.

Согласно обзору [11], методы диффузионной аугментации можно систематизировать по цели преобразования: семантическое редактирование, персонализация и адаптация, а также прикладные сценарии аугментации. При этом в качестве условий генерации могут использоваться текстовые описания [18], изображения [38], метки классов, карты сегментации [39], структурные карты, а также сочетания этих условий за счёт мультимодальности модели.

#### 2.2.1 Методы диффузионной аугментации, управляемые текстовыми инструкциями

К этой группе относятся подходы, в которых генерация или редактирование изображений задаётся текстовой инструкцией, описывающей класс, стиль, домен или желаемое преобразование изображения. Одно из первых эффективных решений, не требующих дообучения модели, было предложено в работе [40]. В этой работе новые изображения

создаются методом редактирования зашумленных представлений исходных образцов с простым фиксированным описанием «a photo of a {класс}». Метод обеспечивает устойчивый прирост качества на задачах с четко разделимыми классами, однако менее эффективен на наборах, где различия между классами выражены слабее.

Важный вклад в развитие методов с текстовым управлением внес подход ALIA (Automated Language-guided Image Augmentation) [22]. Авторы используют предобученные модели компьютерного зрения и обработки естественного языка для генерации описаний доменов набора данных на естественном языке. Процесс состоит из трёх этапов: (1) генерация подробных подписей к каждому изображению исходного набора, (2) обобщение этих подписей большой языковой моделью в короткий список высокоуровневых доменных описаний (например, «птица на фоне леса в пасмурную погоду» или «яркая тропическая растительность»), и (3) использование полученных описаний как текстовых инструкций для редактирования исходных изображений с опорой на текст с помощью модели Stable Diffusion. Метод работает без дообучения диффузионной модели и без ручного подбора текстовых инструкций, обеспечивая при этом значительное семантическое разнообразие при сохранении принадлежности к исходному классу. При этом качество ALIA зависит от точности автоматически сгенерированных подписей и способности языковой модели выделять действительно релевантные доменные факторы.

Ещё одним подходом без дообучения генеративной модели является подход GIF (Guided Imagination Framework) [41]. Данный метод генерирует новые выборки, оптимизируя латентные представления исходных изображений в пространстве признаков предобученной генеративной модели. Для управления семантикой используется набор текстовых подсказок, формируемых по шаблону со случайным выбором доменных и описательных компонентов. Ключевая особенность метода – направленная оптимизация латентных признаков с использованием модели обучения без явной разметки CLIP (Contrastive Language-Image Pre-Training) [32] для оценки согласованности с классом, информативности и разнообразия генерируемых образцов. Синтетические данные объединяются с исходными. Такой подход обеспечивает контролируемое расширение выборки за счёт сохранения семантики классов при одновременном увеличении информативного и структурного разнообразия примеров. Основное ограничение GIF связано с зависимостью от качества CLIP-представлений: CLIP-оценка не всегда надёжно отражает тонкие межклассовые различия.

Дальнейшее развитие связано с адаптацией текстового условия при замороженной диффузионной модели. В методе DA-Fusion [23] для каждого класса обучается специальный классовый псевдотокен с использованием механизма текстовой инверсии: оптимизируется новое векторное представление токена в пространстве текстового энкодера, так чтобы этот токен кодировал визуальные особенности соответствующего класса. Затем выполняется редактирование изображений с инструкцией «a photo of a <псевдо-токен>», а синтетические образцы включаются в набор данных через локальную случайную замену исходных данных. Метод эффективен при малом числе примеров на класс, но его качество ограничено выразительностью одного псевдотокена, который может неполно описывать внутриклассовое разнообразие. Аналогичный принцип лежит в основе метода аугментации DiffuseMix [42], который подразумевает редактирование изображений по текстовым инструкциям с отбором подсказок для стилевых преобразований и последующее смешивание исходного изображения, сгенерированного образца и случайно выбранного фрактального изображения, завершая процесс полной заменой исходных данных синтетическими. Метод хорошо сохраняет семантику в задачах с несколькими объектами, а также при высокой детализации классифицируемых объектов. Однако смешивание исходного изображения, синтетического образца и случайного фрактального изображения может вносить визуальные артефакты и особенности распределения, нехарактерные для целевых данных.

Ещё одним методом является Diff-II [43], который после дообучения диффузионной модели выполняет инверсию изображений (поиск латентного представления, из которого модель





может реконструировать исходное изображение) методом DDIM (Denoising Diffusion Implicit Models). Затем производится сферическая интерполяция латентных представлений пар изображений одного класса, то есть плавный переход между двумя латентными векторами по дуге на гиперсфере (Spherical Linear Interpolation, SLERP). Далее следует двухстадийное восстановление с использованием инструкций, дополненных контекстными суффиксами. Метод сочетает высокую точность воспроизведения класса и значительное разнообразие при малом числе примеров, дисбалансе классов и выходе за пределы обучающего распределения. Ограничением метода Diff-II является чувствительность к качеству латентной интерполяции: она может размывать на изображениях дискриминативные признаки класса.

Отдельного внимания заслуживает подход AGA (Attribute-Guided Augmentation) [44], в котором сегментация используется для выделения и сохранения исходного объекта, а большая языковая модель формирует разнообразные описания фона, по которым диффузионная модель генерирует новые фоновые изображения. Затем исходный объект совмещается с синтетическим фоном, что позволяет повышать разнообразие контекста без существенного искажения признаков класса. Метод ориентирован на классификацию изображений и был успешно применён, в частности, на наборах данных CUB-200-2011 и iWildCam, демонстрируя улучшение качества как на данных, соответствующих обучающему распределению, так и при доменном сдвиге.

### 2.2.2 Методы диффузионной аугментации, управляемые изображениями

Ко второй группе относятся методы, которые отказываются от доминирующей роли текста в пользу прямого использования визуальных признаков, локальных фрагментов или структурных условий. Классическим примером такого управления с изображением в качестве условия служит метод IP-Adapter [38], который позволяет использовать реальные изображения как условие в существующих визуально-языковых моделях. В работе [45] исследуется эффективность модели Stable Diffusion с интеграцией метода IP-Adapter для генерации синтетических изображений растений, демонстрируя, что такой подход существенно расширяет обучающую выборку и повышает точность многоклассовой детекции. Другой подход, DIPSY [36], предлагает генерацию синтетических данных для задач классификации по малому числу примеров. Метод использует двойное задание визуального условия через IP-Adapter: изображение целевого класса задаёт положительное условие, а изображение похожего, но другого класса – отрицательное условие. Расширенная схема бесклассификаторного управления генерацией (classifier-free guidance) [29] позволяет независимо управлять этими условиями, формируя более дискриминативные синтетические образцы. Метод DIPSY продемонстрировал конкурентоспособные результаты на нескольких тестовых наборах.

Также активно развиваются решения на основе подхода ControlNet [46-47]. Так, в работе [48] структурные карты (границы, сегментация, позы) направляют процесс генерации совместно с текстовыми подсказками, которые могут быть автоматически изменены для увеличения разнообразия синтезируемых данных.

Систематический анализ всех этих подходов представлен в обзоре [11], где подчеркивается, что методы с текстовым управлением обеспечивают гибкий семантический контроль, но требуют тщательной подготовки инструкций, в то время как методы, управляемые изображениями, лучше сохраняют визуальную точность и структуру исходных данных, особенно в задачах с несколькими объектами и при смене домена. Однако сильная привязка к изображению-условию может снижать степень семантической вариативности и усиливать уже существующие корреляции между объектом, фоном и доменом съёмки.

Таким образом, существующие диффузионные методы аугментации значительно расширяют возможности компенсации дисбаланса классов и дефицита данных. В настоящей работе на наборах CUB-200-2011 и iWildCam проводится прямое сравнение редактирования

существующих изображений по текстовым инструкциям с помощью моделей Instruct-Pix2Pix и Qwen-Image-Edit с другими диффузионными стратегиями аугментации, включая методы Real Guidance [40], GIF [41], DA-Fusion [23] и другие. Это позволяет объективно оценить преимущества и недостатки контролируемого редактирования уже имеющихся изображений по сравнению с генерацией «с нуля» и визуально-управляемыми стратегиями.

## *3. Методология и эксперименты*

В данном разделе представлено описание решаемой задачи, а также методов и подходов к её решению. Приведена информация о наборах данных, используемых в экспериментах, архитектуре базового классификатора, реализации конвейеров редактирования изображений, стратегии селективного отбора примеров для аугментации, типах редактирующих текстовых инструкций, методах фильтрации синтетических изображений, а также о процедуре проведения экспериментов и оценивания результатов.

### 3.1 Описание решаемой задачи

Рассматривается задача повышения качества классификации в условиях дисбаланса классов за счет генерации дополнительных обучающих примеров методом управляемого редактирования изображений. Пусть задана исходная обучающая выборка

$$D_{train} = \{(x_i, y_i)\}_{i=1}^{N},$$

где $x_i$ – изображение, а $y_i$ – метка класса. Требуется построить расширенную выборку

$$\tilde{D}_{train} = D_{train} \cup D_{aug},$$

где множество синтетических примеров $D_{aug}$ формируется путем редактирования части исходных изображений с помощью текстовых инструкций при сохранении их исходной классовой принадлежности.

В отличие от стандартных аугментаций, предлагаемый подход вносит семантические изменения (фон, ракурс, композиция), сохраняя ключевые признаки класса. При этом ключевое требование к синтетически сгенерированным данным состоит в том, чтобы они расширяли внутриклассовое разнообразие, не разрушая дискриминативные признаки, определяющие принадлежность объекта к исходному классу.

Поскольку стоимость генерации синтетических изображений велика, задача формулируется не только как задача аугментации, но и как задача рационального распределения ограниченного бюджета генерации. Для этого в работе исследуются стратегии селективного отбора исходных примеров для редактирования, учитывающие дисбаланс классов и структуру обучающей выборки. В результате оценивается вклад пяти факторов: модели редактирования, типа текстовой инструкции, стратегии выбора исходных изображений, доли аугментации относительно объёма исходной обучающей выборки, а также наличия и типа фильтрации синтетической добавки к данным.

Критерием эффективности является изменение качества итогового классификатора, обученного на расширенной выборке, по сравнению с базовым режимом обучения без генеративной аугментации. Оценивание производится на фиксированной тестовой выборке, не участвующей ни в генерации, ни в обучении.

### 3.2 Описание конвейера оценки эффективности различных аугментаций

Экспериментальный конвейер состоит из трёх последовательно выполняемых этапов: генерации синтетических изображений, опциональной фильтрации синтетической добавки и обучения классификатора на объединенной выборке.





#### 3.2.1 Этап 1: аугментация изображений

На первом этапе по заданной доле аугментации определяется число синтетических изображений, которое необходимо сгенерировать. После этого в соответствии с выбранной стратегией селективного отбора формируется набор исходных изображений, подлежащих редактированию. Затем к каждому изображению из этого набора применяется выбранный image-to-image метод редактирования. Сгенерированные изображения сохраняются вместе с метаданными (класс, исходное изображение, текстовая инструкция).

#### 3.2.2 Этап 2: фильтрация синтетических изображений

После генерации в части экспериментов выполнялась автоматическая фильтрация синтетической добавки к данным без изменения исходных реальных данных. Отбрасывались изображения, потенциально несогласованные с исходной меткой: семантически нерелевантные примеры, изображения с недостаточно выраженным объектом целевого класса, а также образцы, ошибочно или аномально уверенно распознаваемые внешним классификатором. Для этого использовались CLIP-семантическая проверка и confidence-фильтрация; затем классификатор обучался на объединении исходных данных и оставшихся синтетических изображений.

#### 3.2.3 Этап 3: оценка прироста точности классификации

На третьем этапе базовый классификатор обучается на объединении исходных и синтетических данных, сгенерированных при помощи конкретного метода аугментации с конкретными параметрами. Валидация и тестирование выполняются на неизменных выборках. Фиксируются метрики качества и параметры эксперимента.

Важным свойством реализованного конвейера является его модульность: один и тот же протокол оценивания используется как для редактирования по текстовым инструкциям, так и для альтернативных методов аугментации.

#### 3.2.4 Особенности исследуемого в работе метода Prompt Edit

Метод Prompt Edit, исследуемый в настоящей работе, представляет собой управляемый текстовыми инструкциями подход к аугментации данных на основе визуально-языковых моделей. Он выполняет редактирование уже существующих примеров (image-to-image) с помощью двух диффузионных конвейеров – Instruct-Pix2Pix и Qwen-Image-Edit. Для каждого отобранного изображения случайным образом из заранее сформированного набора выбирается текстовая инструкция, описывающая желаемые семантические изменения (фон, ракурс, освещение или стиль), при строгом сохранении принадлежности объекта к исходному классу.

### 3.3 Наборы данных

Эксперименты проводились на двух наборах данных классификации изображений детализированных объектов. Эти наборы данных имеют различный характер дисбаланса классов.

Первый набор – CUB-200-2011 (Caltech-UCSD Birds-200-2011) [26] – содержит 11 788 изображений 200 видов птиц. Исходно набор сбалансирован, поэтому для моделирования выраженного дисбаланса в обучающей выборке в половине классов (100 классов) количество изображений было случайным образом уменьшено на 80% (в условиях сохранения воспроизводимости экспериментов). Оставшиеся 100 классов сохранили исходный объём. Таким образом, в экспериментальной версии набора наблюдался ярко выраженный дисбаланс, типичный для реальных прикладных задач.

Второй набор – 13-классовое подмножество iWildCam [27] (изображения, полученные с

камер фотоловушек). Этот набор данных изначально обладает естественным дисбалансом классов, характерным для данных экологического мониторинга, где одни виды животных встречаются значительно чаще других. Подмножество было получено путем выбора из набора 13-ти наиболее качественно размеченных классов.

Для каждого набора данных использовалось стандартное разбиение на обучающую, валидационную и тестовую выборки. Тестовая выборка оставалась неизменной и не подвергалась аугментации.

Выбор долей аугментации для разных наборов обусловлен различиями в их размере и степени дисбаланса классов. Набор CUB после искусственного прореживания характеризуется значительным числом классов и выраженным дефицитом примеров в половине из них, что требует более агрессивного синтетического расширения для достижения выравнивания распределения. В связи с этим для него рассматриваются более высокие уровни аугментации – 100% и 250%, позволяющие существенно увеличить представленность редких классов и приблизиться к эффекту насыщения. В случае iWildCam исходный набор изображений обладает естественным дисбалансом, но при этом содержит большее количество изображений на класс, что снижает необходимость чрезмерного увеличения выборки. Кроме того, высокая стоимость генерации и риск накопления шумовых артефактов при больших долях аугментации делают более оправданным использование умеренных значений – 50% и 100%, достаточных для оценки эффективности методов без избыточного роста вычислительных затрат.

## 3.4 Базовый классификатор и условия обучения

В качестве базовой модели классификации в экспериментах использовалась архитектура ResNet-50 в силу ее относительно оптимальной сбалансированности между точностью классификации и вычислительной эффективностью. В разделе 4.8 показано, что выводы, которые следуют из экспериментов с ResNet-50, хорошо переносятся на другие (более современные и в том числе более вычислительно затратные) архитектуры классификаторов.

Модель обучалась с нуля на расширенном обучающем наборе данных (исходные изображения + синтетические образцы). Оптимизатор – SGD со значением параметра momentum = 0.9, параметры learning rate, weight decay и количество эпох были эмпирически подобраны для каждого набора данных отдельно (см. табл. 1).

*Табл. 1. Гиперпараметры обучения классификатора ResNet-50 на наборах данных CUB и iWildCam.*
*Table 1. Hyperparameters for training a classifier on the CUB and iWildCam datasets.*

| Стратегия аугментации | CUB | iWildCam |
|---|---|---|
| Количество эпох | 100 | 50 |
| learning rate | 0.01 | 0.0003 |
| weight decay | 0.001 | 0.001 |

В качестве функции потерь использовалась взвешенная по классам кросс-энтропия, что обеспечило высокую корреляцию между достоверностью (accuracy) и сбалансированной достоверностью (balanced accuracy) (см. табл. 2). Поэтому в качестве основной (целевой) метрики была выбрана достоверность (accuracy) – с учётом её сильной связи со сбалансированной версией. Наблюдаемые в табл. 2 отклонения корреляции от единицы обусловлены незначительными различиями в значениях метрик для некоторых методов диффузионной аугментации, которые в некоторых случаях лежат в пределах относительной погрешности измерений.

Ещё одной важной метрикой классификации, рассмотренной в данном исследовании, является макроусредненная F1-мера (macro F1). Было установлено, что данная метрика демонстрирует высокую корреляцию (см. табл. 2) с общей точностью (accuracy).





*Табл. 2. Значения корреляции метрик при рассмотрении совокупности различных диффузионных аугментаций для наборов данных CUB и iWildCam (архитектура классификатора - ResNet-50).*
*Table 2. Correlation coefficients of classification metrics when considering a set of different diffusion augmentations for the CUB and iWildCam datasets (classifier architecture: ResNet-50).*

| **Наборы данных** | | CUB | | iWildCam | |
|---|---|---|---|---|---|
| **Тип корреляции** | | Корреляция Пирсона | Ранговая корреляция (Спирмена) | Корреляция Пирсона | Ранговая корреляция (Спирмена) |
| **Accuracy & balanced accuracy** | **ResNet-50** | 1.00 | 1.00 | 0.90 | 0.81 |
| | **ViT-L/16** | 1.00 | 1.00 | 0.96 | 0.71 |
| **Accuracy & F1-macro** | **ResNet-50** | 1.00 | 1.00 | 0.92 | 0.95 |
| | **ViT-L/16** | 1.00 | 1.00 | 0.99 | 0.98 |

Следовательно, результаты и выводы, полученные на основе метрики достоверности (accuracy), вполне могут быть применимы и к макроусредненной F1-мере.

## 3.5 Конвейеры редактирования изображений

Для генерации синтетических образцов применялись два современных конвейера визуально-языкового редактирования на базе диффузионных моделей:

1. **Instruct-Pix2Pix** [18] – метод редактирования изображений по текстовым инструкциям, основанный на дообученной модели Stable Diffusion. Конвейер принимает исходное изображение и текстовую инструкцию, описывающую желаемые изменения.
2. **Qwen-Image-Edit** [19] – реализация редактирования на базе мультимодального диффузионного трансформера из семейства Qwen (Qwen-Image-Edit). Метод поддерживает совместное использование изображения и текстовой инструкции и демонстрирует высокую фотореалистичность, а также точность следования инструкциям.

Оба конвейера работают в режиме генерации изображения из изображения и позволяют вносить контролируемые семантические изменения, сохраняя принадлежность объекта к исходному классу.

## 3.6 Стратегии селективного отбора изображений для аугментации

Для выбора изображений, подлежащих редактированию, были разработаны и протестированы пять стратегий, обладающих разной степенью адаптивности к структуре дисбаланса:

1. **Случайная (Random)** – полностью случайный выбор изображений из исходного набора с возможными повторениями.
2. **Полное объединение со случайным дополнением (Full_plus_random)** – при доле аугментации более 100% сначала вся обучающая выборка добавляется целое число раз, а оставшаяся часть добирается случайно без повторений.
3. **Взвешенная по классам (Class_weighted)** – вероятность выбора изображения обратно пропорциональна размеру его класса (изображения из малочисленных классов выбираются чаще).
4. **Равномерно-взвешенная по классам (Uniform_class_weighted)** – для каждого класса заранее рассчитывается квота обратно пропорционально размеру класса, после чего изображения внутри каждого класса выбираются равномерно случайным

образом (равномерно с возможными повторами).

5. **Жадное добавление (Greedy_adding)** – итеративный выбор изображений из наименее представленных классов (внутри каждого класса выбираются равномерно случайным образом) до постепенного выравнивания распределения.

### 3.7 Редактирующие текстовые инструкции и гиперпараметры

Исследовались три типа редактирующих инструкций, сформулированных в максимально унифицированной (доменонезависимой) форме: (1) изменение фона, (2) изменение положения и ракурса объекта и (3) комбинированные изменения (фон + положение + стиль).

Для модели Qwen-Image-Edit [19] дополнительно варьировался гиперпараметр числа шагов вывода (num_inference_steps) в диапазоне от 5 до 50; оптимальное значение 10–20 использовалось в основных экспериментах.

Доля аугментируемых изображений варьировалась от 10% до 250% от исходного размера обучающей выборки, аугментированные изображения добавлялись к исходной обучающей выборке.

### 3.8 Сравнение эффективности различных аугментаций

Сравнение различных аугментаций проводилось в условиях строго фиксированного протокола обучения и оценивания. Для каждого набора данных сохранялись неизменными архитектура классификатора, схема разбиения данных, функция потерь, оптимизатор, правило изменения скорости обучения и набор метрик. Варьировались только параметры, непосредственно относящиеся к аугментации: метод аугментации, модель редактирования, стратегия отбора изображений для генерации, тип текстовых инструкций и доля синтетического расширения обучающей выборки. Фильтрация синтетических изображений исследована отдельно и не применялась при сравнении методов аугментации. Такой дизайн эксперимента позволяет корректно связать наблюдаемые изменения качества именно с исследуемым механизмом генеративной аугментации.

Сопоставление выполнялось относительно базового режима (baseline) без генерации синтетических данных (в условиях «чистого» набора данных с искусственным (CUB) или естественным (iWildCam) дисбалансом классов), а также между различными конфигурациями редактирования. Для каждой конфигурации фиксировались значения метрик accuracy, balanced accuracy и F1-macro, однако в силу их эмпирической высокой корреляции (см. табл. 2) рассматривалась именно первая из этих метрик. Дополнительно регистрировались параметры генерации и обучения, что позволяло анализировать не только абсолютное качество, но и чувствительность результата к выбору стратегии отбора, типу текстовой инструкции и объему аугментации.

## *4. Результаты*

В данном разделе представлены результаты проведённых экспериментов. Основной метрикой качества служила точность классификации (accuracy) на фиксированной тестовой выборке. Для удобства сравнения в таблицах и на диаграммах указывается прирост точности относительно базового обучения без аугментации (baseline).

### 4.1 Сравнение моделей редактирования

Сравнение конвейеров (моделей) Instruct-Pix2Pix [18] и Qwen-Image-Edit [19] при оптимальных настройках (стратегия Greedy_adding, комбинированные текстовые инструкции, доля аугментации 100% для CUB и 50% для iWildCam) показало превосходство Qwen-Image-Edit (см. табл. 3). На наборе CUB модель Qwen обеспечила прирост точности в среднем от +8.6 п.п. до +9.7 п.п. (в зависимости от числа шагов вывода Qwen), в то время как





Instruct-Pix2Pix – лишь +1.7 п.п. На iWildCam соответствующие значения составили +3.6 п.п. и +1.3 п.п. Таким образом, Qwen-Image-Edit превосходит Instruct-Pix2Pix по эффективности семантической аугментации, что объясняется лучшей фотореалистичностью генерируемых изображений (см. рис. 1) и более точным следованием текстовым инструкциям благодаря архитектуре мультимодального диффузионного трансформера.

*Табл. 3. Результаты сравнения моделей редактирования изображений применительно к аугментации наборов данных CUB и iWildCam.*
*Table 3. Results of a comparison of image editing models applied to augmentation of the CUB and iWildCam datasets.*

| Стратегия аугментации | CUB: Accuracy, % | iWildCam: Accuracy, % |
|---|---|---|
| baseline | 55.94 | 80.41 |
| Instruct-Pix2Pix | 57.66 | 81.68 |
| Qwen-Image-Edit | 64.54 | 84.01 |

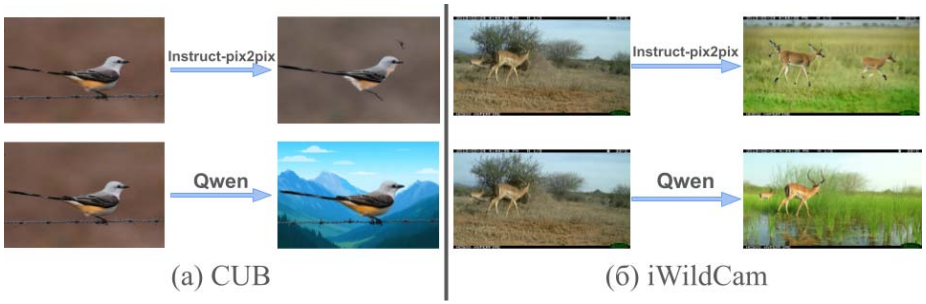



*Рис. 1. Пример редактирования изображений из наборов данных (a) CUB и (b) iWildCam моделями Instruct-Pix2Pix и Qwen-Image-Edit в условиях одинаковой текстовой инструкции.*
*Fig. 1. An example of image editing from (a) CUB and (b) iWildCam datasets using the Instruct-Pix2Pix and Qwen-Image-Edit models, given the same text instruction.*

При формировании примеров на рис. 1 использовались следующие текстовые инструкции:

a) CUB: change the background to a coastal environment, and change the Western Gull position so that it is viewed at a 45-degree angle.
b) iWildCam: without changing the image style and the impala, replace the background with a freshwater wetland marsh and zoom in on the impala.

Отметим, что архитектура Qwen требует значительных ресурсов графических ускорителей при сравнительно медленной скорости редактирования (генерации) изображений.

### 4.2 Влияние стратегий селективного отбора изображений

Серьезное влияние на качество аугментации в условиях дисбаланса оказывает стратегия отбора изображений. Был проведен сравнительный анализ эффективности селективных стратегий выбора изображений для аугментации на обрезанном наборе CUB_200_2011 (см. рис. 2). Доля аугментированных изображений составила 100% от начального размера обучающей выборки обрезанного набора данных, для редактирования использовалась модель Qwen.

На наборе CUB (рис. 2) стратегии Uniform_class_weighted и Greedy_adding показали наилучшие результаты, обеспечивая наиболее адаптивное выравнивание распределения классов. Стратегия Random уступила им примерно на 5.5 процентных пункта при той же доле аугментации. Class_weighted показало промежуточный результат, её отставание от стратегий

Uniform_class_weighted и Greedy_adding объясняется ее нестабильностью при больших долях аугментации из-за переполнения малочисленных классов.

Дополнительно на наборе iWildCam было проведено сравнение стратегий Uniform_class_weighted и Greedy_adding, поскольку структура созданного на CUB дисбаланса не позволяет выявить существенную разницу между ними. В результате Greedy_adding вновь оказалась наиболее эффективной (см. табл. 4) благодаря тому, что она итеративно фокусируется на наименее представленных классах.

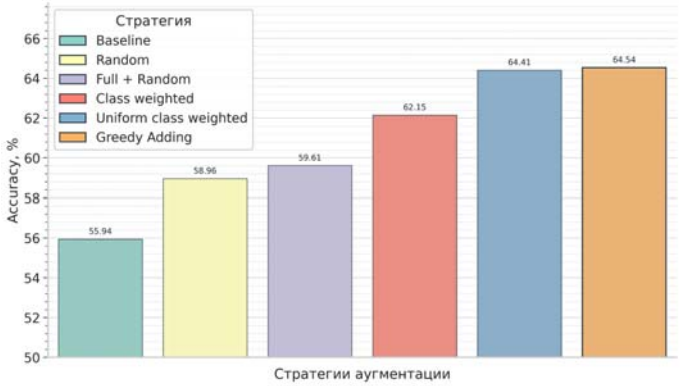



*Рис. 2. Сравнение эффективности селективных стратегий выбора изображений для набора CUB.*
*Fig. 2. Comparison of effectiveness of selective image selection strategies for the CUB dataset.*

*Табл. 4. Результаты применения стратегий Uniform_class_weighted и Greedy_adding для набора данных iWildCam.*
*Table 4. Results of applying the Uniform_class_weighted and Greedy_adding strategies to the iWildCam dataset.*

| Стратегия аугментации | Accuracy, % |
|---|---|
| baseline | 80.41 |
| Uniform_class_weighted | 83.39 |
| Greedy_adding | 84.01 |

## 4.3 Влияние доли аугментации

Зависимость точности классификации от доли аугментируемых изображений при использовании модели Qwen-Image-Edit для набора CUB представлена на рис. 3. Для стратегии Greedy_adding рост достоверности (accuracy) носит практически монотонный характер; значение стабилизируется (в пределах погрешности эксперимента) уже при 200% аугментации на CUB. Стратегия Random также достигает насыщения примерно после того же значения – 200%. Превышение этой величины (200% для обеих стратегий) приводит лишь к минимальному дополнительному приросту точности при одновременном росте вычислительных затрат.

Представленный на рис. 3 график наглядно демонстрирует важность стратегии выбора изображений для аугментации: при использовании Greedy_adding при доле аугментации 25% достигается результат, которого стратегия Random достигает только при доле аугментации около 250%.

## 4.4 Влияние типов редактирующих текстовых инструкций

Выбор типа содержания текстовых инструкций существенно влияет на финальную точность





(см. рис. 4). В условиях применения модели Qwen на наборе CUB инструкции, изменяющие фон (background), оказались более эффективными и обеспечили прирост на 4.1 процентных пункта выше, чем инструкции, изменяющие только положение объекта (см. Приложение).

Комбинированные текстовые инструкции (фон + положение + стиль), составленные как совокупность попарных объединений «фоновых» и «позиционных» инструкций, показали наилучшие результаты, превосходя отдельные типы примерно на 1.9 п.п. Основываясь на результатах этого эксперимента, в последующих экспериментах данной работы используется именно набор комбинированных инструкций (120 шт.), из которого для каждого аугментируемого изображения случайным образом выбирается одна текстовая инструкция.

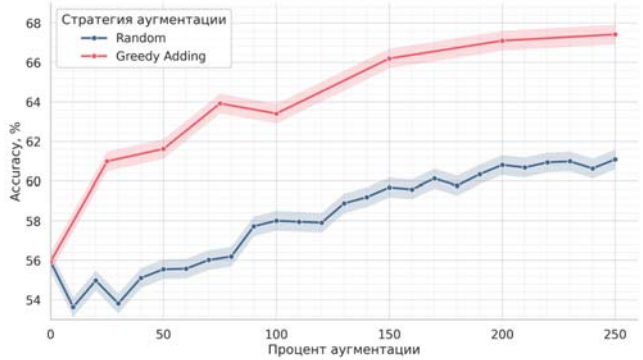



*Рис. 3. График точности классификации (метрика accuracy) при использовании разного размера аугментации набора данных CUB.*
*Fig. 3. Classification accuracy plot when using different augmentation levels on the CUB dataset.*

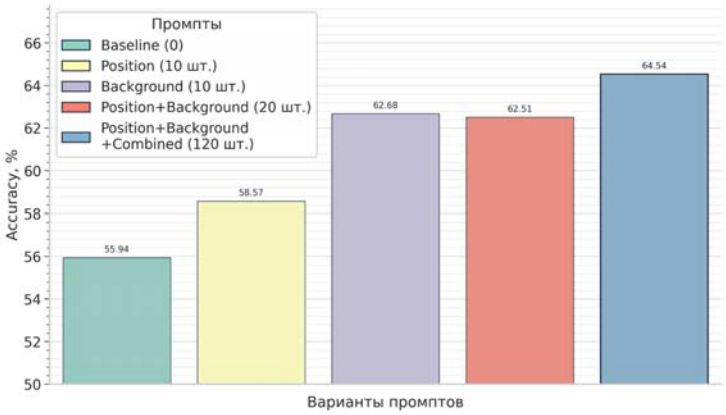



*Рис. 4. Сравнение эффективности применения редактирующих текстовых инструкций различных типов для набора данных CUB.*
*Fig. 4. Comparison of effectiveness of different types of text-based editing instructions for the CUB dataset.*

## 4.5 Влияние числа шагов вывода модели Qwen-Image-Edit

Для Qwen-Image-Edit критически важным гиперпараметром является число шагов вывода (num_inference_steps). Его увеличение приводит к повышению качества генерируемых

изображений, однако также приводит к возрастанию времени редактирования (генерации) изображений. Рис. 5 наглядно демонстрирует, что существенное улучшение качества генерируемых изображений достигается уже при 10 шагах. Дальнейшее увеличение до 20 шагов даёт дополнительный прирост около 0.8 п.п., после чего кривая выходит на плато. Значения выше 30 шагов нецелесообразны, поскольку время генерации растёт линейно, а прирост достоверности (accuracy) становится статистически незначимым.

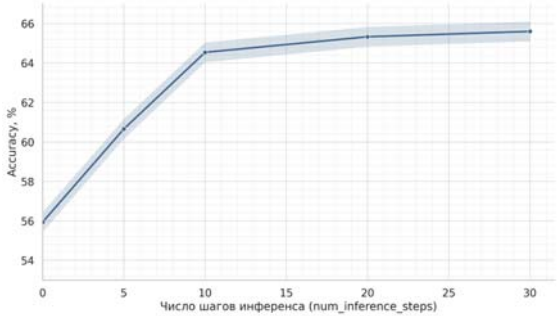



*Рис. 5. График точности классификации (accuracy) в зависимости от количества шагов инференса модели Qwen при аугментации набора данных CUB.*
*Fig. 5. A graph showing classification accuracy as a function of the number of inference steps for the Qwen model when augmenting the CUB dataset.*

## 4.6 Сравнение с традиционными методами аугментации

С целью выяснения сравнительной эффективности исследуемого метода на фоне стандартных подходов к аугментации [49-54] были проведены эксперименты по аугментации наборов данных CUB и iWildCam. Использовалась стратегия выбора изображений Greedy_adding при различных долях аугментации (см. рис. 6 и рис. 7). Стандартные методы аугментации, требующие адаптации под конкретный набор, использовались с политиками аугментации, полученными для ImageNet, без перенастройки под целевые наборы данных.

Предложенный VLM-подход (Qwen-Image-Edit) значительно превосходит традиционные методы аугментации. На рис. 7 показано, что на наборе CUB при доле аугментации 100% VLM-метод обеспечил прирост +8.6 процентных пунктов, в то время как Auto Augment [49], Rand Augment [50] и Trivial Augment [51] дали +4.3 п.п., +6.0 п.п. и +6.3 п.п. соответственно. CutMix [52] и MixUp [53] показали +1.4 п.п. и +3.0 п.п., Random Erasing [54] – +0.4 п.п. При увеличении доли до 250% разрыв только увеличивается: VLM (Prompt Edit Qwen) – +11.4 п.п., в то время как традиционные методы дали не более +7.8 п.п. Одним из объяснений увеличения разницы является большее разнообразие синтетических изображений, генерируемых VLM, по сравнению со стандартными подходами. Однако, в данных условиях эксперимента свою эффективность проявило только редактирование при помощи модели Qwen, в то время как исследуемый метод Prompt Edit с моделью Instruct-Pix2Pix [18] не смог превзойти стандартные аугментации.

На iWildCam при доле аугментированных изображений 50% метод Prompt Edit дал +3.6 п.п., традиционные методы – от -1.3 п.п. (Random Erasing [54]) до +1.1 п.п. (Trivial Augment [51]). Сравнительные диаграммы приведены на рис. 6. Низкую эффективность традиционных методов для набора iWildCam можно объяснить тем, что изображения с фиксированных фотоловушек имеют не слишком большую вариативность ракурса и композиции кадра, поэтому геометрические преобразования могут вносить неестественные для тестовых данных искажения, тогда как семантические изменения, которые вносит метод Prompt Edit, оказываются более релевантными. Таким образом, VLM-редактирование обеспечивает в





среднем больший прирост, чем лучшие традиционные методы, благодаря существенно более выраженным семантическим изменениям.

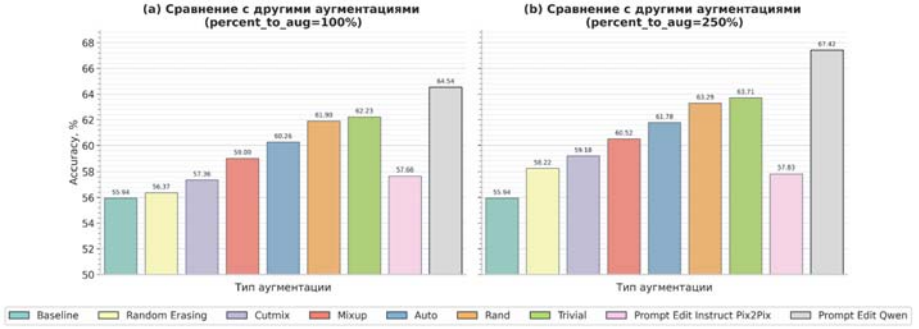



*Рис. 6. Сравнение эффективности исследуемого метода с традиционными аугментациями для набора данных CUB при доле количества аугментированных изображений от размера исходного обучающего набора, равной (a) 100% и (b) 250%.*
*Fig. 6. Comparison of effectiveness of the proposed method with traditional augmentation techniques on the CUB dataset, with the proportion of augmented images relative to the size of the original training set set to (a) 100% and (b) 250%.*

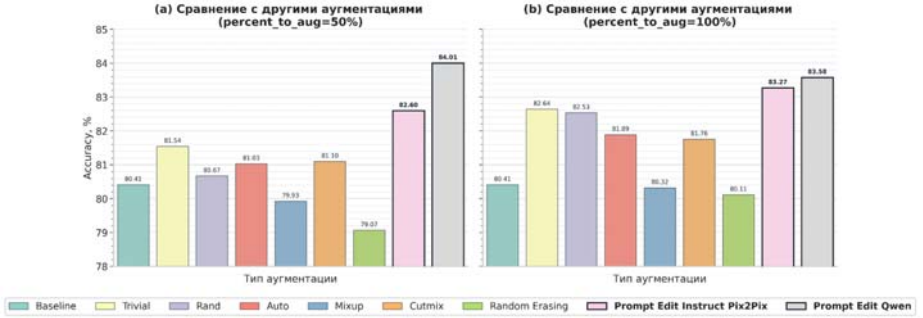



*Рис. 7. Сравнение эффективности исследуемого метода с традиционными аугментациями для набора данных iWildCam при доле количества аугментированных изображений от размера исходного обучающего набора, равной (a) 50% и (b) 100%.*
*Fig. 7. Comparison of effectiveness of the proposed method with traditional augmentation techniques on the iWildCam dataset, with the proportion of augmented images relative to the size of the original training set set to (a) 50% and (b) 100%.*

## 4.7 Сравнение с другими диффузионными методами аугментации

Исследуемый подход Prompt Edit (Qwen) демонстрирует превосходство над другими известными диффузионными стратегиями аугментации. На наборе CUB при использовании стратегии Greedy adding и доле аугментации 100% метод на основе Qwen достигает наивысшей точности среди всех рассмотренных диффузионных методов, уверенно опережая GIF [41], DA-Fusion [23], Diff-Aug [55] и Real Guidance [40]. Это отчасти объясняется качеством самого конвейера Qwen-Image-Edit, который, кроме прочего, задействует при генерации больше временных и вычислительных ресурсов.

На подмножестве iWildCam при доле аугментации 50% метод Prompt Edit (Qwen) показывает результаты, сопоставимые с лучшим из конкурентов (GIF [41]), и превосходит Real Guidance

[40], Diff-Aug и Diff-Mix [55], а также DA-Fusion [23]. При использовании в методе Prompt Edit конвейера Instruct-Pix2Pix получаются сопоставимые с другими диффузионными аугментациями результаты.

Сравнительные диаграммы приведены на рис. 8. Таким образом, эксперименты подтверждают высокую эффективность исследуемого подхода как на искусственно созданном, так и на естественном дисбалансе классов, демонстрируя его конкурентоспособность при сравнении с современными диффузионными методами аугментации.

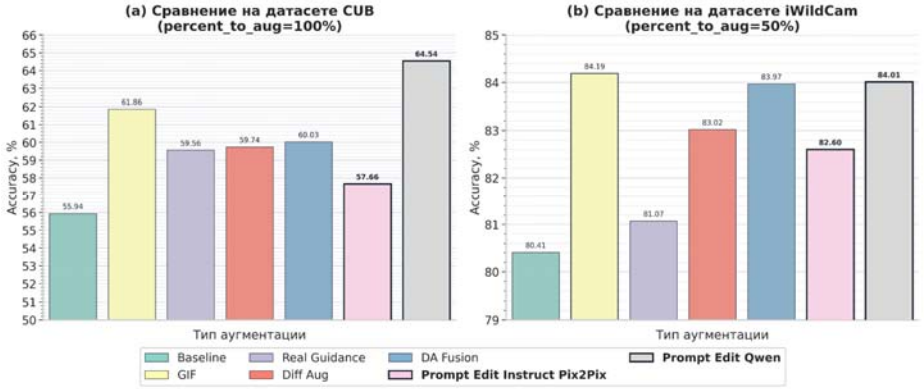



*Рис. 8. Сравнение эффективности исследуемого метода с другими диффузионными аугментациями с использованием классификатора ResNet-50 для наборов данных (a) CUB и (b) iWildCam при доле количества аугментированных изображений от размера исходного обучающего набора, равной (a) 100% и (b) 50%.*

*Fig. 8. Comparison of effectiveness of the proposed method with other diffusion augmentation techniques using the ResNet-50 classifier on (a) CUB and (b) iWildCam datasets, with the proportion of augmented images relative to the size of the original training set set to (a) 100% and (b) 50%.*

## 4.8 Другие архитектуры классификаторов

В силу вариативности методов классификации изображений целесообразно исследовать переносимость выводов, сделанных на основе модели ResNet-50, на другие, в том числе более современные архитектуры. В качестве альтернативной современной архитектуры рассмотрена ViT-L/16 (Vision Transformer Large с размером патча 16×16 пикселей). Для ViT были проведены эксперименты по сравнению различных диффузионных методов аугментации (аналогично разделу 4.7). В целях экономии ресурсов для ViT-L/16 было взято сравнительно небольшое число эпох при обучении (для CUB – 20, для iWildCam – 30), которого оказалось достаточно для успешной сходимости модели. Остальные значения гиперпараметров были аналогичны экспериментам с ResNet-50 (см. табл. 1).

Сравнительный анализ диаграмм на рис. 8 и рис. 9 (ResNet-50 и ViT-L/16 соответственно) показывает, что в силу высокой степени совпадения относительной высоты соответствующих столбцов качественные выводы, которые следуют из раздела 4.7, справедливы также для других архитектур классификаторов, что показано на примере ViT-L/16.

Переносимость выводов также подтверждается высокими значениями корреляции метрик (см. табл. 5), наблюдаемыми при совместном анализе различных методов диффузионной аугментации как единой статистической совокупности.





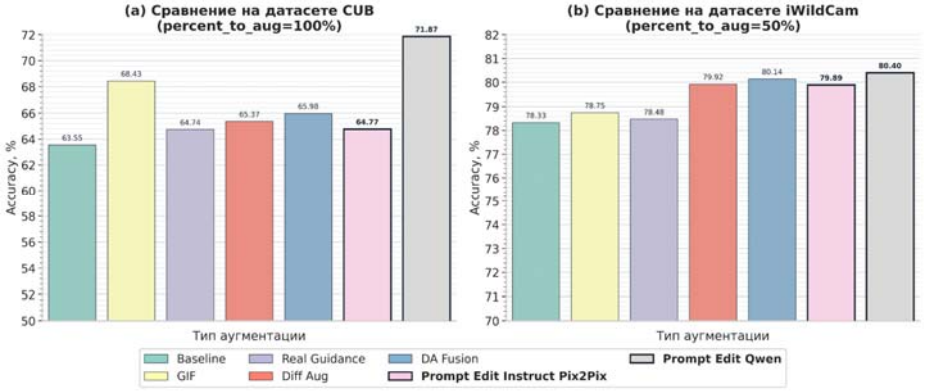



*Рис. 9. Сравнение эффективности исследуемого метода с другими диффузионными аугментациями с использованием классификатора ViT-L/16 для наборов данных (a) CUB и (b) iWildCam при доле количества аугментированных изображений от размера исходного обучающего набора, равной (a) 100% и (b) 50%.*

*Fig. 9. Comparison of effectiveness of the proposed method with other diffusion augmentation techniques using the ViT-L/16 classifier on (a) CUB and (b) iWildCam datasets, with the proportion of augmented images relative to the size of the original training set set to (a) 100% and (b) 50%.*

*Табл. 5. Корреляция метрик классификаторов ResNet-50 и ViT-L/16, обученных на наборах данных, аугментированных различными диффузионными методами.*

*Table 5. Correlation of metrics for ResNet-50 and ViT-L/16 classifiers trained on datasets augmented using various diffusion methods.*

| **Наборы данных** | CUB | | iWildCam | |
|---|---|---|---|---|
| **Тип корреляции** | Корреляция Пирсона | Ранговая корреляция (Спирмена) | Корреляция Пирсона | Ранговая корреляция (Спирмена) |
| **Accuracy** | 0.99 | 0.98 | 0.81 | 0.67 |
| **Balanced accuracy** | 0.99 | 0.98 | 0.83 | 0.81 |
| **f1-macro** | 0.99 | 0.98 | 0.85 | 0.67 |

## 4.9 Фильтрация синтетических изображений

Одним из ограничений генеративной аугментации является возможность появления синтетических изображений, которые формально наследуют исходную метку, но содержат визуальные артефакты, недостаточно выраженный объект целевого класса или семантическое несоответствие исходной категории. Кроме того, часть синтетических изображений может оказаться слишком простой для классификатора: такие примеры корректно распознаются с аномально высокой уверенностью и потенциально вносят меньший вклад в увеличение полезного внутриклассового разнообразия. Поэтому дополнительно была исследована автоматическая фильтрация синтетической добавки к данным перед обучением классификатора. Во всех экспериментах фильтрация применялась только к сгенерированным изображениям; исходные реальные данные не удалялись.

В табл. 6 приведено сравнение режима без фильтрации и пяти вариантов автоматического отбора синтетических изображений. Фильтры группы «Уверенность» используют внешний классификатор, который предсказывает класс каждого синтетического изображения и степень уверенности в этом предсказании. Если классификатор с высокой уверенностью относит изображение не к исходному классу, такой пример считается потенциально

«ошибочным» и удаляется. Также при некоторых конфигурациях фильтрации дополнительно удаляются «простые» примеры – изображения, которые классификатор относит к правильному классу с аномально высокой уверенностью. В вариантах «Уверенность (val): ошибочные и простые» и «Уверенность (train): ошибочные и простые» различается только выбор данных, по которым оцениваются пороги уверенности: валидационная выборка или исходная обучающая выборка соответственно. Вариант «Уверенность (synthetic train): ошибочные» реализует поиск вероятных нарушений метки средствами библиотеки cleanlab: на основе матрицы предсказанных вероятностей выявляются синтетические примеры, для которых исходная метка выглядит несогласованной с предсказаниями классификатора. Наконец, фильтр «CLIP-семантика» проверяет соответствие изображения целевому домену с помощью CLIP [32]: синтетический пример сравнивается с положительным текстовым описанием и набором отрицательных описаний, после чего удаляются изображения, более близкие к нерелевантным описаниям.

*Табл. 6. Влияние фильтрации синтетической добавки к данным на точность (accuracy) классификации, %.*
*Table 6. Effect of filtering the synthetic subset on accuracy, %.*

| **Набор данных** | CUB | | iWildCam | |
|---|---|---|---|---|
| **Модель редактирования** | Instruct-Pix2Pix | Qwen-Image-Edit | Instruct-Pix2Pix | Qwen-Image-Edit |
| **Без фильтрации** | **57.93** | 64.54 | 81.92 | 83.51 |
| **Уверенность (val): ошибочные и простые** | 55.09 | 62.70 | **82.54** | 83.25 |
| **Уверенность (val): ошибочные** | 57.29 | 64.56 | 81.69 | 83.51 |
| **Уверенность (train): ошибочные и простые** | 57.45 | 64.42 | 81.79 | 83.78 |
| **Уверенность (synthetic train): ошибочные** | 57.25 | 63.68 | 81.96 | **83.91** |
| **CLIP-семантика** | 57.65 | **65.27** | 82.02 | 83.63 |

Результаты показывают, что фильтрация синтетических изображений не является универсально улучшающей процедурой: её эффективность зависит как от набора данных, так и от модели редактирования. На наборе CUB для изображений, сгенерированных Instruct-Pix2Pix, все варианты фильтрации приводят к снижению достоверности (accuracy) относительно режима без фильтрации. Это означает, что автоматическая отбраковка может удалять не только некорректные, но и полезные разнообразные примеры, тем самым уменьшая регуляризующий эффект аугментации.

Для Qwen-Image-Edit на CUB наилучший результат даёт фильтр «CLIP-семантика»: достоверности (accuracy) возрастает с 64.54% до 65.27%. Это указывает на то, что для более качественной генеративной модели полезной оказывается мягкая семантическая проверка, удаляющая преимущественно явно нерелевантные изображения. Напротив, фильтрация по уверенности с удалением «простых» примеров снижает качество до 62.70%, что говорит о чрезмерной агрессивности такого критерия в задаче классификации птиц.

На iWildCam эффект фильтрации менее выражен, но в ряде случаев положителен. Для Instruct-Pix2Pix наилучший результат достигается при фильтре «Уверенность (val): ошибочные и простые»: достверность (accuracy) увеличивается с 81.92% до 82.54%. Для Qwen-Image-Edit максимальное значение достверности (accuracy) получено при фильтрации «Уверенность (synthetic train): ошибочные» – 83.91% против 83.51% без фильтрации. При этом величина прироста остаётся умеренной, поэтому фильтрация в данном случае не является определяющей.





Отдельно следует отметить различие между наборами данных. На CUB удаление «простых» корректно классифицируемых синтетических изображений заметно ухудшает результат: для Instruct-Pix2Pix достоверность (accuracy) снижается с 57.93% до 55.09%, а для Qwen-Image-Edit – с 64.54% до 62.70% при использовании валидационных порогов. Вероятно, в условиях выраженного искусственного дисбаланса такие изображения не являются избыточными: они могут усиливать представление редких классов и повышать устойчивость классификатора. На iWildCam, напротив, удаление «простых» примеров в отдельных случаях оказывается полезным, что может быть связано с естественным характером дисбаланса и большей неоднородностью изображений, полученных с фотоловушек.

Таким образом, фильтрация синтетической добавки может дополнительно улучшать качество VLM-аугментации, однако её вклад существенно меньше, чем вклад выбора модели редактирования, стратегии отбора изображений и типа текстовых инструкций. Наиболее практически оправданными оказались семантическая CLIP-фильтрация для Qwen-Image-Edit на CUB и фильтрация вероятных ошибочных синтетических примеров для Qwen-Image-Edit на iWildCam. Агрессивные confidence-критерии, особенно с удалением корректных «простых» примеров, следует применять осторожно, поскольку они могут уменьшать полезное внутриклассовое разнообразие синтетической выборки.

### 4.10 Обсуждение полученных результатов

Полученные результаты подтверждают высокую эффективность предложенного метода Prompt Edit для аугментации данных в условиях дисбаланса классов. Наиболее значимый прирост точности обеспечивается сочетанием мощной мультимодальной модели Qwen-Image-Edit с адаптивными стратегиями отбора (Greedy_adding и Uniform_class_weighted) и комбинированными текстовыми инструкциями, затрагивающими как фон, так и положение объекта. Преимущество VLM-редактирования над традиционными методами (AutoAugment, RandAugment, CutMix, MixUp) достигает 3.6 п.п. на CUB и 2.5 п.п. на iWildCam, что объясняется существенно большим семантическим разнообразием синтетических образцов по сравнению с низкоуровневыми трансформациями.

Сравнение с другими диффузионными стратегиями (GIF, DA-Fusion, Diff-Aug, Real Guidance) показывает, что Prompt Edit на базе Qwen уверенно превосходит конкурентов на CUB и демонстрирует сопоставимые или лучшие результаты на iWildCam. Эффективность Prompt Edit в первую очередь связана с качеством редактирования Qwen-Image-Edit, обеспечивающего фотореалистичные и согласованные с инструкциями изображения, а также с адаптивным отбором исходных примеров и использованием комбинированных инструкций, повышающих внутриклассовое разнообразие. Полученные данные свидетельствуют, что предложенный подход представляет собой эффективный и практически применимый инструмент компенсации дисбаланса классов, хотя и требует заметных вычислительных ресурсов при использовании модели Qwen.

## *5. Заключение*

Настоящая работа продемонстрировала, что управляемое VLM-редактирование изображений на базе диффузионных моделей (в частности, Instruct-Pix2Pix [18] и Qwen-Image-Edit [19]) является эффективным инструментом компенсации дисбаланса классов в задачах классификации изображений. Предложенный подход позволяет вносить контролируемые семантические изменения (фон, ракурс, композиция и стиль). За счёт этого он обеспечивает более высокое внутриклассовое разнообразие по сравнению с традиционными геометрическими и фотометрическими аугментациями, а также демонстрирует преимущество над рядом существующих диффузионных методов (Real Guidance [40], GIF [41], DA-Fusion [23] и др.).

Эксперименты на двух реальных наборах данных – CUB-200-2011 [26] с искусственно индуцированным дисбалансом и 13-классовом подмножестве iWildCam [27] с естественным дисбалансом – подтвердили превосходство VLM-подхода. При оптимальной конфигурации (Qwen-Image-Edit + стратегия Greedy_adding + комбинированные текстовые инструкции) достигнут прирост достоверности (accuracy) на 8.6 процентных пунктов (CUB, 100% аугментации) и 3.6 п.п. (iWildCam, 50% аугментации) относительно baseline, что значительно превышает результаты стандартных аугментаций и некоторых конкурентных диффузионных методов.

Таким образом, работа вносит четыре основных вклада:

1. Систематическое сравнение VLM-редактирования (Prompt Edit) с широким спектром альтернативных методов аугментации, демонстрирующее конкурентоспособность метода и превосходство такой аугментации с использованием Qwen-Image-Edit над традиционными аугментациями;
2. Разработка и валидация пяти стратегий селективного отбора, в разной степени адаптированных к структуре дисбаланса классов;
3. Количественная оценка влияния типа текстовых инструкций и гиперпараметров на эффективность VLM-подхода к аугментации изображений.
4. Оценка влияния автоматической фильтрации синтетических изображений на качество классификации для разных наборов и моделей редактирования.

Несмотря на достигнутые результаты, существуют ограничения. Во-первых, Qwen-Image-Edit демонстрирует высокую эффективность, но требует значительных вычислительных ресурсов. Во-вторых, качество синтетических изображений всё ещё зависит от точности следования инструкциям и может содержать редкие артефакты, потенциально влияющие на обучение. В-третьих, эксперименты ограничены двумя наборами данных с детализированными объектами, однако планируется дальнейшее расширение исследований на другие наборы и домены данных.

Перспективы дальнейших исследований включают:

1. Масштабирование подхода на более крупные и сложные наборы данных (медицинские и спутниковые изображения, аэрофотосъемка);
2. Разработку гибридных конвейеров, объединяющих VLM-редактирование с методами, управляемыми изображениями (ControlNet, IP-Adapter);
3. Исследование влияния VLM-аугментации на устойчивость моделей к сдвигу распределения;
4. Более детальное исследование zero-shot и few-shot сценариев в задаче классификации изображений.

Предложенный метод открывает путь к созданию более устойчивых и обобщающих классификаторов в условиях реальных ограниченных и несбалансированных данных, что особенно актуально для автономного вождения, экологического мониторинга и медицинской диагностики, а также для многих других прикладных задач.

## Список литературы / References

[1]. Avetisyan A., Tigranyan S., Asatryan A., Mashkova O., Skorik S., Ananev V., Markin Y. Deep neural networks generalization and fine-tuning for 12-lead ECG classification. Biomedical Signal Processing and Control, 2024, vol. 93, 106160.

[2]. Paheding S., Saleem A., Siddiqui M.F.H., Rawashdeh N., Essa A., Reyes A.A. Advancing horizons in remote sensing: a comprehensive survey of deep learning models and applications in image classification and beyond. Neural Computing and Applications, 2024, vol. 36(27), pp. 16727-16767.





[3]. Либерман А.И. Решение задачи классификации объектов на изображении методами компьютерного зрения в различных сферах человеческой деятельности. Вестник ВГУ. Серия: Системный анализ и информационные технологии, 2024, т. 3, с. 74-91.

[4]. Altalhan M., Algarni A., Alouane M.T.H. Imbalanced data problem in machine learning: A review. IEEE access. 2025, vol. 13, pp. 13686-13699.

[5]. Lim J.Y., Lim K.M., Lee C.P., Tan Y.X. A review of few-shot image classification: Approaches, datasets and research trends. Neurocomputing, 2025, vol. 649, 130774.

[6]. Liu Y., Bai Y., Che X., He J. Few-shot fine-grained image classification: A survey. In 2022 4th International Conference on Natural Language Processing (ICNLP). IEEE, March 2022, pp. 201-211.

[7]. Yang S., Wu S., Liu T., Xu M. Bridging the gap between few-shot and many-shot learning via distribution calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, vol. 44, no. 12, pp. 9830-9843.

[8]. Cheung T.H., Yeung D.Y. A survey of automated data augmentation for image classification: Learning to compose, mix, and generate. IEEE transactions on neural networks and learning systems, 2023, vol. 35, no. 10, pp. 13185-13205.

[9]. Chen Y., Yan Z., Zhu Y. A comprehensive survey for generative data augmentation. Neurocomputing, 2024, vol. 600, 128167.

[10]. Yang S., Xiao W., Zhang M., Guo S., Zhao J., Shen F. Image data augmentation for deep learning: A survey. arXiv preprint 2022. DOI: 10.48550/arXiv.2204.08610.

[11]. Alimisis P., Mademlis I., Radoglou-Grammatikis P., Sarigiannidis P., Papadopoulos G.T. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions. Artificial Intelligence Review, 2025, vol. 58, no. 4, p. 112.

[12]. Kirichenko P., Ibrahim M., Balestriero R., Bouchacourt D., Vedantam S.R., Firooz H., Wilson A. G. Understanding the detrimental class-level effects of data augmentation. Advances in Neural Information Processing Systems, 2023, vol. 36, pp. 17498-17526.

[13]. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems. 2020, vol. 33, pp. 6840-6851.

[14]. Song J., Meng C., Ermon S. Denoising diffusion implicit models. arXiv preprint 2020. DOI: 10.48550/arXiv.2010.02502.

[15]. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684-10695.

[16]. Huang Y., Huang J., Liu Y., Yan M., Lv J., Liu J., Xiong W., Zhang H., Cao L., Chen S. Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, vol. 47, no. 6, pp. 4409–4437.

[17]. Zhang C., Zhang C., Zhang M., Kweon I.S., Kim J. Text-to-image diffusion models in generative AI: A survey. arXiv preprint 2023. DOI: 10.48550/arXiv.2303.07909.

[18]. Brooks T., Holynski A., Efros A.A. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18392-18402.

[19]. Wu C. et al. Qwen-image technical report. arXiv preprint, 2025. DOI: 10.48550/arXiv.2508.02324.

[20]. Wu Z., Kolkin N., Brandt J., Zhang R., Shechtman E. Turboedit: Instant text-based image editing. In European Conference on Computer Vision. Cham: Springer Nature Switzerland, September, 2024, pp. 365-381.

[21]. Mokady R., Hertz A., Aberman K., Pritch Y., Cohen-Or D. Null-text inversion for editing real images using guided diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 6038-6047.

[22]. Dunlap L., Umino A., Zhang H., Yang J., Gonzalez J.E., Darrell T. Diversify your vision datasets with automatic diffusion-based augmentation. Advances in neural information processing systems, 2023, vol. 36, 79024-79034.

[23]. Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. Effective data augmentation with diffusion models. arXiv preprint 2023. DOI: 10.48550/arXiv.2302.07944.

[24]. Li Z., Shi Y., Gao Y., Xu D. Diffusion-Based Data Augmentation for Image Recognition: A Systematic Analysis and Evaluation. International Journal of Computer Vision, 2026, vol. 134, no. 3, p. 126.

[25]. Çağlar Ü.M., Temizel A. Grounding Synthetic Data Generation With Vision and Language Models. arXiv preprint 2026. DOI: 10.48550/arXiv.2603.09625.

[26]. Wah C., Branson S., Welinder P., Perona P., Belongie S. The caltech-ucsd birds-200-2011 dataset, 2011.

[27]. Beery S., Agarwal A., Cole E., Birodkar V. The iwildcam 2021 competition dataset. arXiv preprint 2021. DOI: 10.48550/arXiv.2105.03494.
[28]. Goodfellow I. et al. Generative adversarial networks. Communications of the ACM, 2020, vol. 63, no. 11, pp. 139-144.
[29]. Ho J., Salimans T. Classifier-free diffusion guidance arXiv preprint 2022. DOI: 10.48550/arXiv.2207.12598.
[30]. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, October 2015, pp. 234-241.
[31]. Oktay O. et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint 2018. DOI: 10.48550/arXiv.1804.03999.
[32]. Radford A. et al. Learning transferable visual models from natural language supervision. In International conference on machine learning. PmLR, July 2021, pp. 8748-8763.
[33]. Esser P., Kulal S., Blattmann A., Entezari R., Müller J., Saini H., Levi Y., Lorenz D., Sauer A., Boesel F., Podell D., Dockhorn T., English Z., Rombach R. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the 41st International Conference on Machine Learning, 2024, vol. 235, pp. 12606–12633.
[34]. Bai S. et al. Qwen3-vl technical report. arXiv preprint 2025. DOI: 10.48550/arXiv.2511.21631.
[35]. Kingma D.P., Welling M. Auto-encoding variational bayes. arXiv preprint 2013. DOI: 10.48550/arXiv.1312.6114.
[36]. Boudier L., Manganelli L., Tsonis E., Dufour N., Kalogeiton V. Training-free synthetic data generation with dual IP-adapter guidance. arXiv preprint. 2025. DOI: 10.48550/arXiv.2509.22635.
[37]. Lingenberg T., Reuter M., Sudhakaran G., Gojny D., Roth S., Schaub-Meyer S. Diagen: Semantically diverse image augmentation with generative models for few-shot learning. In DAGM German Conference on Pattern Recognition. Cham: Springer Nature Switzerland, September, 2024, pp. 148-163.
[38]. Ye H., Zhang J., Liu S., Han X., Yang W. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint 2023. DOI: 10.48550/arXiv.2308.06721.
[39]. Couairon G., Careil M., Cord M., Lathuiliere S., Verbeek J. Zero-shot spatial layout conditioning for text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2174-2183.
[40]. He R. et al. Is synthetic data from generative models ready for image recognition? arXiv preprint 2022. DOI: 10.48550/arXiv.2210.07574.
[41]. Zhang Y., Zhou D., Hooi B., Wang K., Feng J. Expanding small-scale datasets with guided imagination. Advances in neural information processing systems, 2023, vol. 36, pp. 76558-76618.
[42]. Islam K., Zaheer M.Z., Mahmood A., Nandakumar K. Diffusemix: Label-preserving data augmentation with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27621-27630.
[43]. Wang Y., Chen L. Inversion circle interpolation: Diffusion-based image augmentation for data-scarce classification. In Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 25560-25569.
[44]. Rahat F., Hossain M.S., Ahmed M.R., Jha S.K., Ewetz R. Data augmentation for image classification using generative AI. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, February, 2025, pp. 4173-4182.
[45]. Deng B., Lu Y. Weed image augmentation by IP-adapter-based stable diffusion for multiclass weed detection. In Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications III. SPIE, May 2025, vol. 13459, pp. 82-96.
[46]. Deng B., Lu Y. Weed image augmentation by ControlNet-added stable diffusion for multi-class weed detection. Computers and Electronics in Agriculture, 2025, vol. 232, 110123.
[47]. Zhang L., Rao A., Agrawala M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836-3847.
[48]. Benkedadra M. et al. Cia: Controllable image augmentation framework based on stable diffusion. In 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, August 2024, pp. 600-606.
[49]. Cubuk E.D., Zoph B., Mane D., Vasudevan V., Le Q.V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 113-123.





[50]. Cubuk E.D., Zoph B., Shlens J., Le Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 702-703.
[51]. Müller S.G., Hutter F. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In Proceedings of the IEEE/CVF international conference on computer vision. 2021, pp. 774-782.
[52]. Yun S., Han D., Oh S.J., Chun S., Choe J., Yoo Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023-6032.
[53]. Zhang H., Cisse M., Dauphin Y. N., Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv preprint 2017. DOI: 10.48550/arXiv.1710.09412.
[54]. Zhong Z., Zheng L., Kang G., Li S., Yang Y. Random erasing data augmentation. In Proceedings of the AAAI conference on artificial intelligence, April 2020, vol. 34, no. 7, pp. 13001-13008.
[55]. Wang Z. et al. Enhance image classification via inter-class image mixup with diffusion model. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 17223-17233.

## *Информация об авторах / Information about authors*

Дмитрий Александрович МАТЮШИН – исследователь отдела информационных систем ИСП РАН; бакалавр МФТИ. Научные интересы: методы доверенного ИИ, аугментация данных, диффузионные модели, визуально-языковые модели, методы генерации и редактирования изображений.

Dmitry Alexandrovich MATIUSHIN – Researcher at the Information Systems Department, ISP RAS; Bachelor's degree from the Moscow Institute of Physics and Technology. Research interests: methods of trustworthy AI, data augmentation, diffusion models, visual-language models, methods for image generation and editing.

Олег Игоревич БЫКОВ – исследователь центра доверенного искусственного интеллекта ИСП РАН, аспирант ИСП РАН. Научные интересы: доверенный искусственный интеллект, диффузионная генерация изображений, методы борьбы с небезопасным контентом (NSFW), объектно-ориентированные разработки (Object-Oriented Design, OOD), мультимодальные модели (Visual Language Mode, VLM).

Oleg Igorevich BYKOV – Researcher at the Center for Trusted Artificial Intelligence, ISP RAS; PhD student at ISP RAS. Scientific interests: trusted AI, diffusion image generation, methods for combating NSFW, OOD, multimodal models (VLM).

Иван Ильич БУБЕНКОВ – исследователь центра доверенного искусственного интеллекта ИСП РАН; бакалавр МГУ. Научные интересы: методы доверенного искусственного интеллекта, методы борьбы с NSFW, алгоритмы генерации изображений.

Ivan Ilyich BUBENKOV – Researcher at the Center for Trusted Artificial Intelligence, ISP RAS; Bachelor's degree from Moscow State University. Scientific interests: trusted AI methods, methods for combating NSFW, image generation algorithms.

# Приложение

## A. Текстовые инструкции для набора данных CUB

«Позиционные» текстовые инструкции (изменяющие позицию объекта классификации на изображении):

1. "change the {} position so that it is viewed from behind",
2. "change the {} position so that it is viewed from the side",
3. "change the {} position so that it is viewed from the front",
4. "change the {} position so that it is viewed at a 45-degree angle",
5. "without changing the {} position, place the {} above the camera viewpoint",
6. "without changing the {} position, place the {} below the camera viewpoint",
7. "change the {} head so that it is looking to the side",
8. "change the {} head so that it is looking upward",
9. "change the {} head so that it is looking downward",
10. "change the {} head so that it is looking directly at the camera"

«Фоновые» текстовые инструкции (изменяющие фон на изображении):

1. "change the background to a forest environment, keep the {} unchanged",
2. "change the background to a mountain environment, keep the {} unchanged",
3. "change the background to a coastal environment, keep the {} unchanged",
4. "change the background to a wetland environment, keep the {} unchanged",
5. "change the background to a grassland environment, keep the {} unchanged",
6. "change the background to a snowy environment, keep the {} unchanged",
7. "change the background to a tropical jungle environment, keep the {} unchanged",
8. "change the background to an urban park environment, keep the {} unchanged",
9. "change the background to a desert environment, keep the {} unchanged",
10. "change the background to a blurred forest environment, keep the {} unchanged"





## B. Текстовые инструкции для набора данных iWildCam

«Позиционные» текстовые инструкции (изменяющие позицию объекта классификации на изображении):

1. "This photo depicts an {}. Without changing the image style, add another {} of the same species to the foreground."
2. "This photo depicts an {}. Without changing the image style, add another {} of the same species to the background."
3. "This photo depicts an {}. Without changing the image style, make the image clearer and the {} more visible."
4. "This photo depicts an {}. Without changing the image style, turn the {} in the other direction."
5. "This photo depicts an {}. Without changing the image style, zoom in on the {}."
6. "This photo depicts an {}. Without changing the image style, turn the {}'s head towards the camera."

«Фоновые» текстовые инструкции (изменяющие фон на изображении):

1. "This photo depicts an {}. Without changing the image style and the {}, replace the background with a dry savanna grassland."
2. "This photo depicts an {}. Without changing the image style and the {}, replace the background with a dense tropical rainforest."
3. "This photo depicts an {}. Without changing the image style and the {}, replace the background with a semi-arid shrub steppe."
4. "This photo depicts an {}. Without changing the image style and the {}, replace the background with a sparse desert grassland."
5. "This photo depicts an {}. Without changing the image style and the {}, replace the background with an open dry forest clearing."
6. "This photo depicts an {}. Without changing the image style and the {}, replace the background with a freshwater wetland marsh."
7. "This photo depicts an {}. Without changing the image style and the {}, replace the background with a dense African forest with many branches."

Комбинированные текстовые инструкции формируются из перечисленных путем их попарного объединения в одну инструкцию, например: "change the background to a forest environment, and change the {} position so that it is viewed from behind" (комбинация первых инструкций из обоих списков для набора CUB). Всего таким образом было использовано 120 инструкций для CUB и 55 инструкций для набора iWildCam (с учетом объединения перечисленных текстовых инструкций с комбинированными).