СРАВНЕНИЕ МЕТОДОВ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫХ СЕТЕЙ И ДИФФУЗИОННЫХ МОДЕЛЕЙ ДЛЯ ВОССТАНОВЛЕНИЯ АУДИОСИГНАЛОВ ПОСЛЕ ПОТЕРЬ ПРИ СЖАТИИ

Д. С. Кирпичев, *Пензенский государственный технологический университет, kirpichev.1999@mail.ru.*

УДК 004.89

Аннотация. В статье представлена сравнительная оценка характеристик генеративно-состязательных сетей и диффузионных моделей при реализации задачи восстановления аудиосигналов после потерь при сжатии. Раскрыты основные характеристики генеративно-состязательных сетей, а также диффузионных моделей. Сравнительная характеристика осуществляется на основе спектрограмм и проводится по таким параметрам, как архитектура, методы обучения, качество, стабильность обучения, скорость генерации, уровень подавления шумов, сложность и т.д. Также определены основные преимущества и недостатки различных моделей.

***Ключевые слова*:** генеративно-состязательные сети; диффузионные модели; восстановление аудиосигнала; нейросети; искусственный интеллект.

COMPARISON OF GENERATIVE ADVERSARIAL NETWORKS (GANS) AND DIFFUSION MODELS FOR LOSSY AUDIO SIGNAL RESTORATION

D.S. Kirpichev, *Penza State Technological University.*

Annotation. The article provides a comparative characteristic of generative adversarial networks and diffusion models in the implementation of the task of restoring audio signals after compression losses. The article provides the main characteristics of generative adversarial networks, as well as diffusion models, comparative characteristics are given by such parameters as: architecture, training methods, quality, stability, training, generation speed, noise suppression level, complexity, etc. Comparative characteristics are given by the main parameters, based on spectrograms, and the main advantages and disadvantages of various models are given.

***Keywords*:** generative adversarial networks; diffusion models; audio signal restoration; neural networks; artificial intelligence.

Введение

Актуальность задачи сравнения генеративно-состязательных сетей (*GAN*) и диффузионных моделей заключается в растущем интересе к генеративным моделям в искусственном интеллекте и их применении в различных областях, таких как искусство, дизайн и промышленность.

Понимание компромиссов между этими моделями помогает выбрать правильный инструмент для конкретных прикладных нужд. Например, *GAN* отличаются скоростью и реалистичностью, а диффузионные модели обеспечивают превосходную детализацию и стабильность обучения.

Также актуальность исследования заключается в возможности улучшить алгоритмы и расширить области применения генеративных моделей. Например, диффузионные модели могут быть адаптированы для широкого спектра задач. Целью настоящей статьи является сравнительный анализ диффузионных моделей и методов генеративно-состязательных сетей в области восстановления аудиосигналов после потерь при сжатии.

Для восстановления аудиосигналов после потерь при сжатии можно использовать следующие методы:

- Избавление от артефактов сжатия. Для этого применяют алгоритмы, которые по рисунку форманта человеческого голоса в низких и средних частотах способны синтезировать, т. е. реконструировать утраченные при сжатии частоты.
- Компенсация входного уровня сигнала. Это делается с помощью параметров *Makeup Gain* или *Output Gain*, которые позволяют восстанавливать входной уровень сигнала, утраченный в ходе компрессии за счет сокращения динамического диапазона.
- Выравнивание по фазе или коррекция. Этот метод помогает исправить возможное «неправильное изображение» в стереополе.
- Временное восстановление. С его помощью можно восстановить слабый звук в начале, например, удара в барабан, который может быть потерян при более низком качестве.
- Использование нейросетей и обучающихся моделей.

Рассмотрим более подробно распространенные и эффективные методики восстановления аудиосигналов после потерь при сжатии.

Архитектура генеративно-состязательных нейросетей

Генеративные антагонистические нейросети (*GAN*) представляют собой класс алгоритмов глубокого обучения, построенных на дуэли двух компонентов: создателя и распознавателя. Первый компонент синтезирует искусственные данные из исходного шумового вектора. Второй выступает в роли классификатора, определяющего подлинность поступающих данных – принадлежат ли они реальному набору или сгенерированы оппонентом [1].

Обучение протекает через последовательное совершенствование обоих модулей. Их соревнование направлено непосредственно на то, чтобы создатель научился производить настолько правдоподобные артефакты, что они смогут обмануть распознавателя. Следует отметить, данная технология достигла передовых рубежей в синтезе изображений и находит применение в иных сферах, включая создание музыки как в нотном представлении, так и в виде аудиосигналов.

Применение *GAN*-архитектур в звуковой генерации имеет давнюю историю. Однако их тренировка сопряжена со сложностями из-за конкурентной природы задач компонентов. Упомянутое является фундаментальным ограничением модели. С целью преодоления упомянутых трудностей и повышения качества/скорости обучения в проекте *Pjloop-GAN* используется концепция *Projected GAN*. Ее суть – интеграция в распознаватель предобученной сети извлечения признаков со стохастическими элементами, препятствующими вырождению модели [2]. На рис. 1 представлена генеративная состязательная сеть (*GAN*) [3, 4].

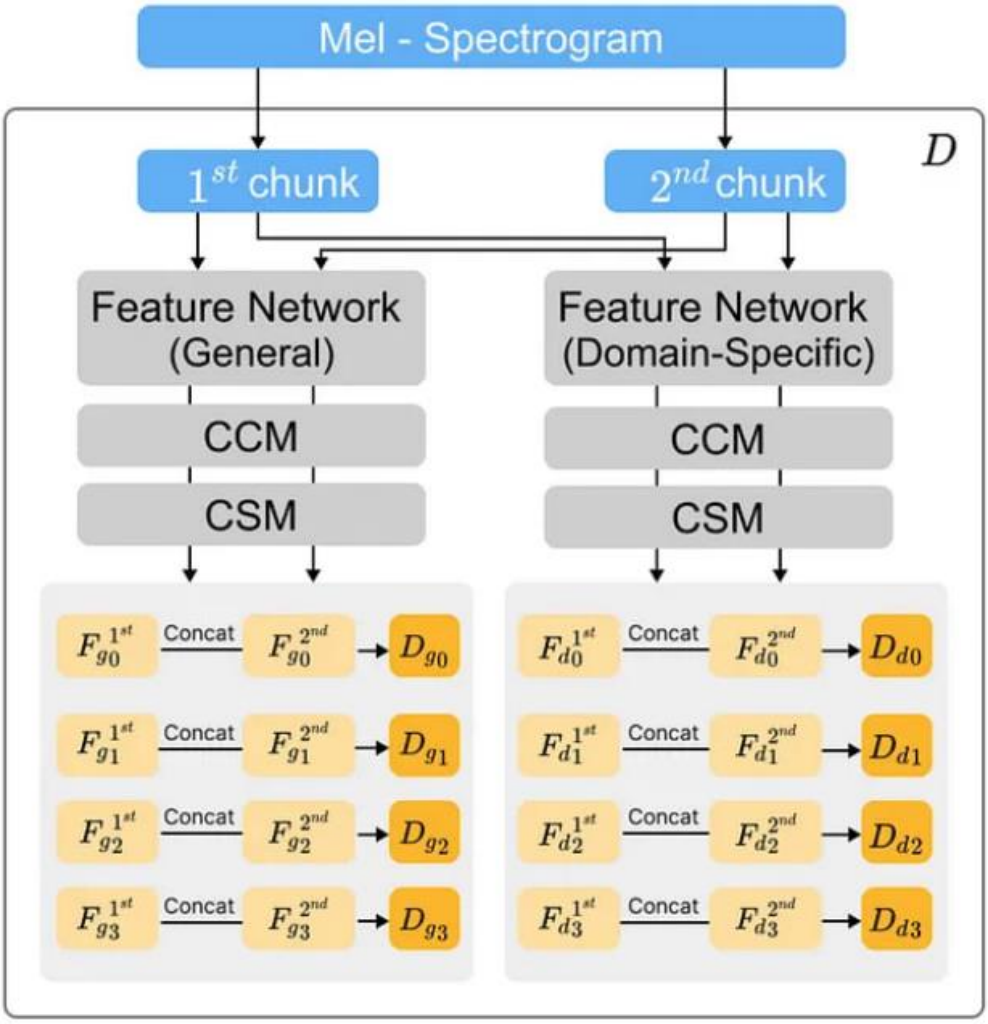

Рисунок 1

Pjloop-GAN оперирует четырьмя «объектными проекторами», перенаправляя анализ с исходных данных на их распределение в многомерном пространстве признаков. Архитектура объединяет всего три ключевые составные части: предобученный экстрактор признаков, модули кросс-канальной модуляции (*CCM*), модули кросс-масштабной модуляции (*CSM*). CCM осуществляет перекрестное взаимодействие признаков по каналам, а *CSM* дополнительно комбинирует их по масштабам. Вносимая обоими модулями стохастичность расширяет способность распознавателя анализировать все пространство признаков. Механизм обновления потерь в *Projected GAN* агрегирует выходные сигналы от всех задействованных классификаторов.

Диффузионные модели, получившие широкое признание в синтезе изображений и реставрации аудио, пока реже используются с целью выполнения первичной генерации звука. Создание реалистичного аудио – сложная задача, требующая учета множества аспектов: временной динамики сигнала, наложения различных звуковых слоев и высокой детализации [5].

Одним из заметных проектов в этой нише, работающих с ограниченными данными, выступает *audio-diffusion-pytorch*. Его принцип генерации аналогичен процессу в диффузионных сетях для изображений: модель *U-Net* последовательно обрабатывает данные с поэтапно снижаемым уровнем шума, инвертируя процесс зашумления с целью создания правдоподобных сэмплов из целевого распределения. Следует отметить, что с целью оптимизации скорости,

реалистичности применяются специальные техники, например, сжатие через автоэнкодер (сокращает временную длительность за счет увеличения каналов) или спектральные преобразования (наращивают канальность, но укорачивают звуковую дорожку). На рис. 2 представлены диффузные нейросети для генерации звуковых файлов [6].

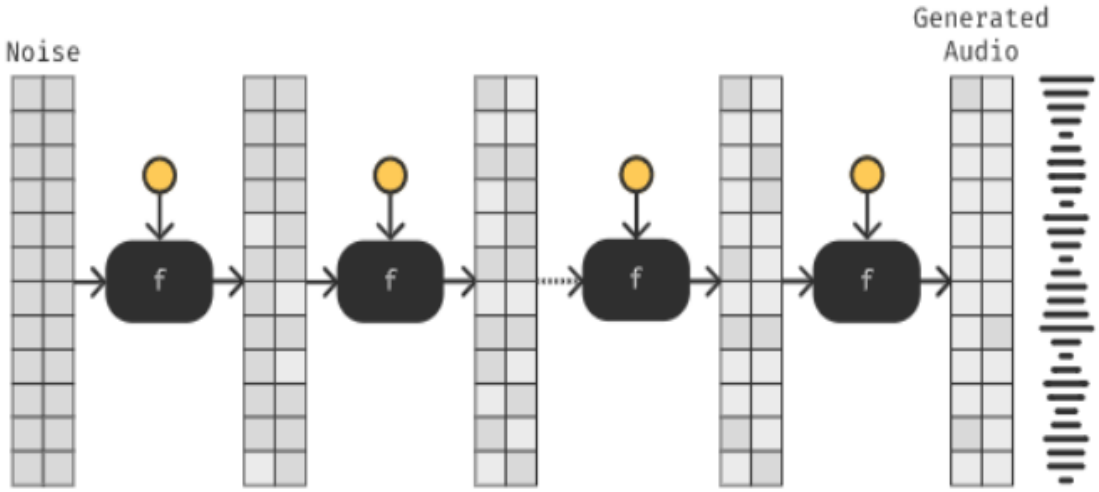

Рисунок 2

Диффузионный вокодер представляет собой генеративную модель, предназначенную для обработки аудиосигналов. Его архитектура предполагает преобразование входного сигнала в спектрограмму, – она затем подвергается пространственному выравниванию посредством транспонированной одномерной свертки. Следует отметить, данная операция учитывает параметры временного окна анализа и количество частотных полос. Далее модель апсемплирует выровненные данные, увеличивая размерность признакового пространства. Полученные многоканальные признаки подаются непосредственно на вход *U-Net* архитектуры, реализующей диффузионный процесс синтеза выходного сигнала с существенно улучшенными акустическими характеристиками [7, 8].

Сравнительный анализ спектрограмм после восстановления

Анализ спектрограмм позволяет сформулировать следующий вывод: оптимизация по критерию минимизации среднеквадратичной ошибки (СКО) провоцирует избыточное сглаживание выходного сигнала модели, следствием чего является подавление высокочастотных составляющих в предсказании. Примечательно, что предсказание, сформированное генеративной моделью, демонстрирует значение СКО, превышающее данный показатель у исходного искаженного сигнала, однако при этом успешно реконструирует речевое содержание. Схожий феномен наблюдается в задачах повышения разрешения изображений (*super-resolution*) [9]. На рис. 3 представлена сравнительная спектрограмма:

А) В исходном виде.

Б) Искаженная после потерь при сжатии.

В) В диффузионной модели обработки.

Г) В обработке *GAN*-сети.

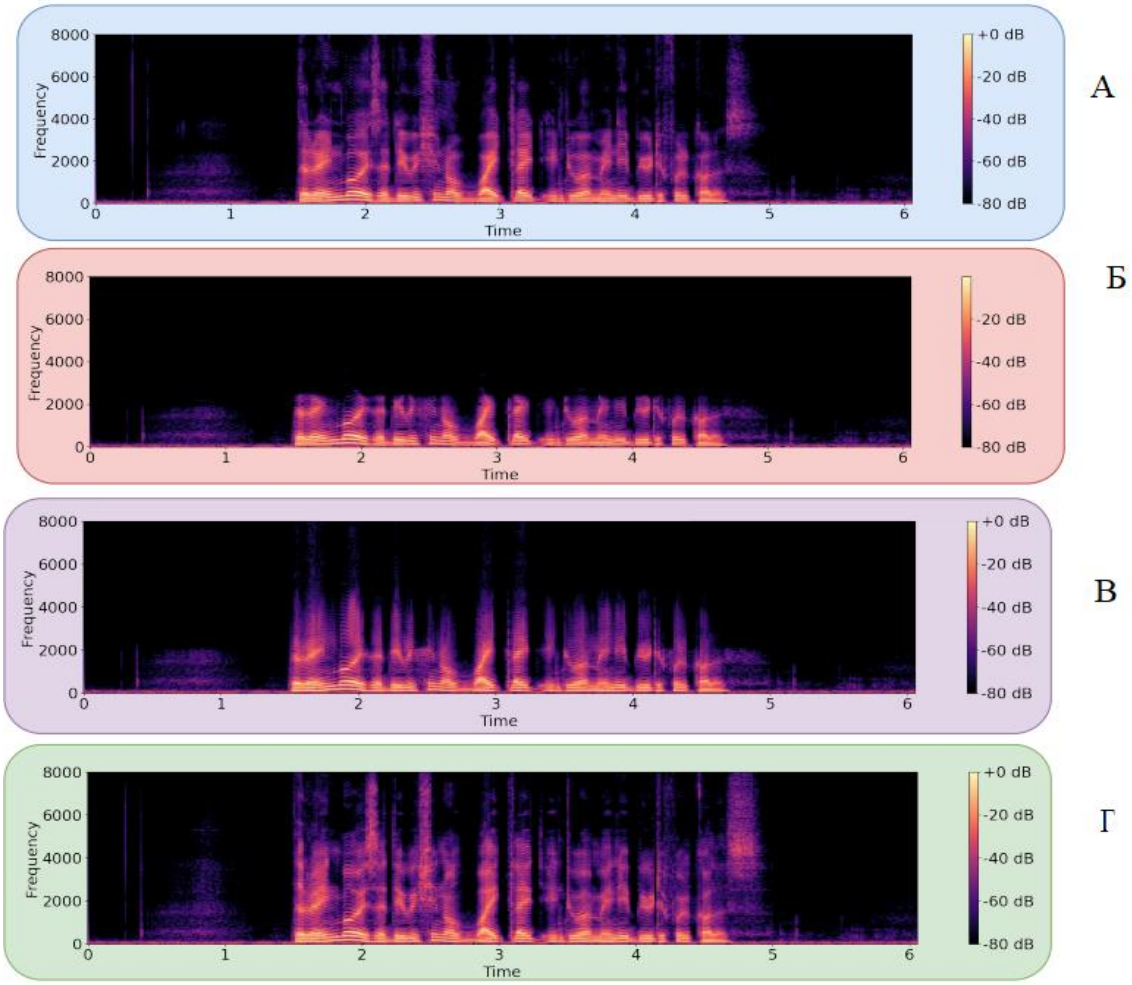

Рисунок 3

Основная практическая задача модели улучшения речи заключается в восстановлении «чистого» аудиосигнала, сохраняющего ключевые атрибуты исходной речевой записи: индивидуальные характеристики голоса, лингвистическую информацию, просодические особенности. Следовательно, модель синтезировать новую речь не должна, а обязана реконструировать существующую, эмулируя ее запись в идеальных студийных условиях. С математической точки зрения это эквивалентно восстановлению наиболее вероятной версии целевого сигнала чистой речи (y) при заданном искаженном входе (x): $y = arg\ max_x\ p\_чистый(y|x)$. Учитывая данную вероятностную постановку задачи, можно утверждать: архитектура обучения, основанная непосредственно на генеративных состязательных сетях (*GAN*), является концептуально соответствующей непосредственно для решения проблемы улучшения качества речи [10].

В рамках сравнительного анализа генеративных моделей (*GAN* и диффузионные подходы) следует отметить [11]:

Сильные стороны *GAN:*

- Исключительная детализация выходных данных: способность генерировать аутентичные аудиообразцы, информацию с высоким уровнем визуального или звукового правдоподобия.
- Широкий спектр применения: успешно адаптируется под разнородные задачи, например, синтез звука, реконструкцию аудиопотоков, аугментацию данных.

Слабые стороны *GAN:*

- Проблемы сходимости: процесс обучения склонен к нестабильности, проявляющейся, например, в схлопывании вариативности (генератор воспроизводит лишь узкий набор сходных образцов).
- Требовательность к конфигурации: необходимость кропотливой оптимизации параметров, структурных элементов сети (с целью достижения работоспособности).

Преимущества диффузионных моделей:

- Высокая точность воспроизведения деталей: эффективно фиксируют тонкие нюансы, сложные паттерны данных благодаря методике последовательного улучшения качества.
- Устойчивость процесса обучения: обучение характеризуется большей предсказуемостью, меньшим числом сбоев режимов (относительно *GAN*).

Недостатки диффузионных моделей:

- Замедленное создание образцов: процедура генерации требует больше времени, чем *GAN* (существует необходимость прохождения множества стадий очистки шума).
- Высокие вычислительные затраты: существенная нагрузка (имеется в виду ресурсная, увеличенное время выполнения обработки, вызванное итеративной природой пошагового удаления шумовых артефактов).

Принимая во внимание требование к академической строгости, а также избегание «машинного» стиля, ниже представлена табл. 1, систематизирующая ключевые характеристики генеративно-состязательных сетей (*GAN*) и диффузионных моделей. В табл. 1 представлены сравнительные характеристики генеративно-состязательных сетей (*GAN*) и диффузионных моделей [12].

Таблица 1.

Аспект	**Генеративно-состязательные сети (*GAN*)**	**Диффузионные модели**
Основная архитектура	Взаимодействие генератора и дискриминатора в рамках состязательного процесса.	Основаны на принципах прямой (зашумление) и обратной (деноизинг) диффузии.
Принцип обучения	Состязательное обучение, использующее минимаксную или вассерштейновскую функцию потерь.	Обучение на задаче шумоподавления, часто с использованием среднеквадратичной ошибки (*MSE*) и многостадийной доработки.
Качество генерируемых данных	Способны создавать высокореалистичные, визуально правдоподобные образцы.	Демонстрируют исключительно высокое разрешение, проработку мельчайших деталей.
Стабильность обучения	Часто характеризуются нестабильностью, склонностью к коллапсу мод, чувствительностью к гиперпараметрам.	В целом более устойчивы и менее подвержены проблемам сходимости (в процессе обучения).

Аспект	Генеративно-состязательные сети (*GAN*)	Диффузионные модели
Вычислительные ресурсы	Могут быть ресурсоемкими, но обеспечивают относительно более высокую скорость инференса.	Обычно требуют значительных вычислительных мощностей, существенного времени обучения и генерации.
Скорость генерации	Относительно высокая за счет прямого, одношагового процесса синтеза.	Сравнительно медленнее из-за итеративного процесса шумоподавления (многократные шаги).
Механизм управления шумом	Минимальная или неявная обработка шума; акцент непосредственно на общую реалистичность выходных данных.	Специализированное обучение на распознавание и эффективное устранение помех на каждом этапе.
Сложность имплементации	Как правило, воспринимаются как относительно более простые в реализации, первичной настройке.	Характеризуются более сложной архитектурой, требуют более изощренных подходов к обучению.
Целевые области применения	Идеально подходят для сценариев реального времени, интерактивных систем; эффективны в работе с аудиоданными.	Оптимальны для высококачественного синтеза изображений и детальной реконструкции; активно исследуются в других модальностях.

Источник: составлено автором на основе [3, 4, 15].

Заключение

В области генеративного моделирования сложилась выраженная дихотомия между двумя ведущими технологиями: генеративно-состязательными сетями (*GAN*) и диффузионными моделями. Ключевое различие между ними заключается в принципиально разном балансе скорости инференса и качестве генерируемых данных [13].

GAN-архитектуры демонстрируют выдающуюся производительность в задачах, требующих в режиме реального времени генерации; их способность к быстрому синтезу визуально правдоподобных данных обусловила их доминирование (в интерактивных системах, динамическом контенте). Однако потенциал нивелируется фундаментальной проблемой нестабильности обучения, проявляющейся в трудностях достижения сходимости, зависимости от инициализации параметров, что усложняет их практическую имплементацию [14].

Напротив, диффузионные модели представляют собой парадигму, ориентированную на достижение максимально возможного качества и фотореализма. Они способны воспроизводить сложнейшие высокочастотные детали, что позволило им установить рекорды качества во многих бенчмарках. Платой за такую детализацию является высокая вычислительная сложность и, как следствие, значительное время инференса, что накладывает существенные ограничения на их использование в чувствительных ко времени приложениях.

Следовательно, практический выбор модели – решение, продиктованное требованиями конкретного сценария. Если приоритетом является оперативная

генерация, предпочтение отдается *GAN*. А если безупречная детализация и стабильность результата ставится во главу угла, то выбираются диффузионные модели. Понимание данного компромисса является необходимым условием для эффективного применения генеративных технологий [15].

Литература

1. Архангельская Е.О., Кадурин А.А., Николенко С.И. Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2022. – 430с.
2. Воробьев Е.Г. Сжатие двоичных кодов на основе традиционных методов и использования псевдорегулярных чисел // Известия СПбГЭТУ «ЛЭТИ», 2015. – № 5. – С. 23-29.
3. Сакулин С.А., Алфимцев А.Н. Защита изображения человека от распознавания нейросетевой системой на основе состязательных примеров // Вестник компьютерных и информационных технологий, 2020. – Т. 17. – № 2 (188). – С. 32-38
4. DeepZip: Lossless Data Compression using Recurrent Neural Networks / Mohit Goyal, Kedar Tatwawadi, Shubham Chandak, Idoia Ochoa. arXiv:1811.08162v1 [cs.CL] 20 Nov 2018. URL: https://arxiv.org/abs/1811.08162
5. Generative Adversarial Networks (GANs) vs Diffusion Models https://neerc.ifmo.ru/wiki/index.php?title=Автокодировщик
6. Image Compression using Backprop. URL: https://web.archive.org/web/20070828112920/http://neuron.eng.wayne.edu/bpImageCompression9PLUS/bp9PLUS.html
7. Lucas Pinheiro Cinelli, et al. Variational Autoencoder // Variational Methods for Machine Learning with Applications to Deep Networks. Springer, 2021. – P. 111-149.
8. Zhihao Duan, Ming Lu, Jack Ma, Yuning Huang, Zhan Ma, Fengqing Zhu. QARV: Quantization-Aware ResNet VAE for Lossy Image Compression // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. – Vol. 46. – P. 436- 450.
9. Generative Adversarial Nets, Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu et al. arXiv:1406.2661v1 [stat.ML] 10 Jun 2014. URL: https://arxiv.org/abs/1406.2661
10. Neural Video Compression using GANs for Detail Synthesis and Propagation / Fabian Mentzer, Eirikur Agustsson, Johannes Ball´e, et al. arXiv:2107.12038v3 [eess.IV] 12 Jul 2022. URL: https://arxiv.org/abs/2107.12038
11. Generative Compression, Shibani Santurkar, David Budden, Nir Shavit. arXiv:1703.01467v2 [cs.CV] 4 Jun 2017. URL: https://arxiv.org/abs/1703.01467.
12. GAN Compression: Efficient Architectures for Interactive Conditional GANs / Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, Song Han // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. – Vol. 44. – P. 9331-9346.
13. Li, Wei; Gauci, Melvin; Gross, Roderich. «A Coevolutionary Approach to Learn Animal Behavior Through Controlled Interaction». Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation Amsterdam, The Netherlands: ACM, 2013. – pp. 223-230.
14. Generative Adversarial Nets (GAN) [Электронный ресурс] – URL: https://neerc.ifmo.ru/wiki/index.php?title=Generative_Adversarial_Nets_(GAN)&mobileaction=toggle_view_desktop
15. Karras T., Aila T., Laine S., Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. URL: https://arxiv.org/abs/1710.10196