

Д. А. Колобова, М. П. Шлеймович

ГЕНЕРАЦИЯ ВИДЕО ПОДСТИЛАЮЩЕЙ ПОВЕРХНОСТИ НА ОСНОВЕ ОДИНОЧНОГО СНИМКА

Ключевые слова: нейронные сети, генерация видео, обработка изображений, кадр, диффузионные модели, интерполяция, цветовые характеристики.

Генерация видео является одной из наиболее актуальных и сложных задач в области искусственного интеллекта и компьютерного зрения. Ее решение открывает широкие возможности для креативных индустрий, бизнеса, образования, маркетинга. Однако генерация продолжительного, семантически связного видео с высоким разрешением остается нерешенной проблемой. Этим объясняется необходимость создания новых, а также исследования уже существующих моделей. В данной статье проводится сравнительный анализ основных методов генерации видео: вариационных автоэнкодеров (VAE), генеративно-состязательных сетей (GAN), авторегрессионных, flow-based и диффузионных моделей. Рассматриваются их ключевые архитектурные особенности, достоинства и недостатки. Особое внимание в работе уделено диффузионным моделям, которые на данный момент являются передовым подходом для решения задачи генерации видео. За последние несколько лет появилось огромное количество диффузионных моделей генерации видео, среди которых наиболее известными являются Sora (OpenAI), Gen-3 (Runway), Kandinsky (Sber AI), Stable Video Diffusion (Stability AI). Однако большинство из них являются закрытыми, коммерческими продуктами, исходный код и архитектура которых недоступны для исследования и модификации. Для реализации генерации в работе используется диффузионная модель с открытым исходным кодом Stable Video Diffusion. Практическая часть исследования включает генерацию видео на основе исходного изображения подстилающей поверхности, а также анализ полученного результата. Сгенерированные видеопоследовательности могут быть использованы для симуляции различных полетных сценариев и расширения датасетов для беспилотных летательных аппаратов (БПЛА). В ходе анализа качества сгенерированного видео выявлено, что для предотвращения накопления артефактов и ошибок генерации требуется дополнительная обработка последовательности кадров после 10-13 кадра видео. Анализ проводился с помощью набора метрик, отражающих изменение цветовых характеристик и текстуры сгенерированного видео.

D. A. Kolobova, M. P. Shleimovich

VIDEO GENERATION OF THE UNDERLYING SURFACE BASED ON A SINGLE IMAGE

Keywords: neural networks, video generation, image processing, frame, diffusion models, interpolation, color characteristics.

Video generation is one of the most urgent and challenging tasks in the field of artificial intelligence and computer vision. Solution of this task opens up wide opportunities for creative industries, business, education, and marketing. However, generating long-lasting, semantically coherent high-resolution video remains an unsolved problem. This explains necessity to create new models, as well as research existing models. This article provides a comparative analysis of the main methods of video generation: variational autoencoders (VAE), generative-adversarial networks (GAN), autoregressive, flow-based and diffusion models. Their key architectural features, advantages and disadvantages are considered. Special attention is paid to diffusion models, which are currently an advanced approach for solving the problem of video generation. Over the past few years, a huge number of diffusion video generation models have appeared, among which the most famous are Sora (OpenAI), Gen-3 (Runway), Kandinsky (Sber AI), Stable Video Diffusion (Stability AI). However, most of them are closed, commercial products, the source code and architecture of which are inaccessible for research and modification. The open source Stable Video Diffusion model is used to implement the generation. The practical part of the study includes video generation based on the original image of the underlying surface, as well as analysis of the result. The generated video sequences can be used to simulate various flight scenarios and expand datasets for unmanned aerial vehicles (UAVs). During the analysis of the quality of the generated video, it was revealed that additional processing of the sequence of frames after 10-13 frames of the video is required to prevent the accumulation of artifacts and generation errors. The analysis was carried out using a set of metrics reflecting changes in the color characteristics and texture of the generated video.

Введение

Генерация видео — это создание видео с помощью алгоритмов искусственного интеллекта на основе текстовых подсказок, изображений, сценариев или даже речи. Она позволяет превращать простые запросы в динамичный видеоконтент.

Генерация позволяет создавать новые данные при недостатке их для обучения моделей. Также она позволяет увеличить их разнообразие, учитывая различные условия освещения, контент, ракурсы для улучшения обобщающей способности моделей.

Как у любой другой задачи, у генерации видео есть свои технические особенности, а именно:

1) Создание видео требует значительно больше вычислительных ресурсов, чем генерация статичных изображений.

2) Искусственный интеллект должен понимать временную согласованность, физику движения и причинно-следственные связи в трехмерном пространстве.

Данные особенности являются причиной сложности подобной задачи, в связи с чем в открытом доступе до сих пор нет модели генерации видео, удовлетво-

ряющей сразу трем показателям: качество видео, скорость генерации, длительность видео. Поэтому основной задачей в области генерации видео на текущий момент является создание продолжительного видео (более минуты) высокого качества (FullHD) при уменьшении вычислительных и временных затрат.

Генерация видео с помощью искусственного интеллекта применяется в самых разных областях — от маркетинга и развлечений до образования и дизайна, например [1]:

1) Кинопроизводство и анимация. Генеративные модели используются для создания спецэффектов, реалистичной анимации персонажей и фоновых сцен;

2) Игровая индустрия. Технология применяется для создания игровых сцен, генерации анимированных персонажей и динамического контента;

3) Социальные сети. В социальных сетях генерация видео помогает создавать контент, привлекающий внимание пользователей;

4) Маркетинг. Компании используют генерацию для быстрого создания персонализированных рекламных видеороликов.

5) Интерактивные видеоуроки. Генерация видео позволяет создавать увлекательные и наглядные обучающие материалы, например, анимированные объяснения сложных понятий.

6) Виртуальные симуляторы. Генерация видео помогает создавать симуляции для обучения, например, моделируя сложные научные эксперименты или рабочие ситуации.

7) Цифровое искусство. Художники и дизайнеры используют генерацию видео для создания уникальных анимированных произведений, экспериментируя с различными стилями и эффектами.

Также в последнее время получила развитие идея создания модели мира (General world model). Это система искусственного интеллекта, которая имеет внутреннее представление об окружающей среде и использует его для моделирования будущих событий в этой среде, т.е. понимает окружающий мир, его физику и способна предсказывать отклик на свои действия или внешние события. Хорошая модель генерации видео для имитации физики реального мира должна его понимать. Поэтому создание моделей генерации видео высокого качества - перспективный путь к созданию универсальных симуляторов физического мира. Модель мира, в свою очередь, позволит обучать автономные транспортные средства, роботов и т.д. действиям в окружающем мире [2,3].

Например, GAIA-1 от Wayve - это генеративная модель мира для автономного вождения. GAIA-1 использует возможности диффузионных моделей для генерации видео - реалистичных сценариев вождения. Таким образом, она может использоваться как продвинутый симулятор, позволяющий генерировать неограниченное количество данных, включая примеры для обучения и проверки систем автономного вождения [4].

Подобные модели также могут быть использованы для обучения беспилотных летательных аппаратов (БПЛА). Разработка таких моделей позволила

бы генерировать реалистичные сценарии полета в различных условиях: городской среде, сложных погодных условиях (дождь, снег, туман), появление препятствий, отказ системы, а также симулировать различные полетные сценарии (доставка, съемка, взлет, посадка, обнаружение объектов и т.д.).

За последние несколько лет появилось огромное количество моделей генерации видео, среди которых наиболее известными являются Sora (OpenAI), Veo (Google), Gen-3 (Runway), Kandinsky (Sber AI), Stable Video Diffusion (Stability AI). Однако большинство из этих моделей являются закрытыми, коммерческими продуктами, исходный код и архитектура которых недоступны для исследования и модификации. В условиях такой закрытости остро ощущается необходимость как создания новых моделей, способных генерировать длительные видео без потери пространственно-временной согласованности, так и усовершенствования существующих открытых подходов. Например, модель Video LDM (Video Latent Diffusion Model) была разработана на базе модели Stable Diffusion и способна генерировать высококачественные видео на основе текстовых запросов с разрешением до 1280×2048 пикселей [5].

Исследование возможностей моделей с открытым исходным кодом, их ограничений и артефактов генерации является основой для будущих работ в области генерации видео — будь то создание новых архитектур или оптимизация существующих. Именно поэтому целью данной работы является исследование методов генерации видео и способов улучшения его качества на основе анализа цветовых и текстурных характеристик. Такой анализ позволяет не только оценить текущее состояние открытых моделей, но и обосновать необходимость применения методов постобработки, направленных на устранение выявленных недостатков и приближение качества открытых решений к уровню их закрытых аналогов.

Обзор существующих методов

Методы генерации данных появились задолго до появления нейронных сетей. К таким методам относятся метод Монте-Карло, марковские цепи и автокорреляционные модели. Однако эти методы имели ограниченные возможности и не могли создавать сложные структуры данных, такие как изображения. Но они стали основой для создания первых генеративных нейронных сетей.

Одной из первых архитектур генеративных нейронных сетей были автокодировщики (AE) и их улучшенная версия — вариационные автокодировщики (VAE) [6]. В архитектуре AE есть три главных компонента: кодировщик, латентное пространство и декодер [7]. Кодировщик переводит входные данные в более компактное пространство, называемое латентным, а декодер восстанавливает их. Изначально, благодаря своей способности представлять данные в пространстве более низкой размерности, их использовали для сжатия/восстановления данных. Ключевая особенность появившихся в 2013 году VAE заключается в том, что он обучается не просто сжимать данные, а изучать вероятностное распределение

этого латентного пространства, что позволяет генерировать новые, разнообразные примеры данных [6]. Недостатком VAE является генерация размытых изображений. Однако их способность уменьшать размерность данных впоследствии была применена в других методах генерации изображений и видео, например, в латентных диффузионных моделях [5].

В 2014 году появилась новая архитектура – генеративные состязательные сети (GAN) [8]. GAN состоит из двух сетей – генератора и дискриминатора, «состязующихся» между собой. Генератор учится создавать реалистичные изображения из случайного шума, а дискриминатор пытается отличить сгенерированные изображения от реальных. В контексте генерации видео генератор преобразует шум в кадры видео, стараясь соблюсти логическую связанность последовательности кадров, а дискриминатор оценивает реалистичность получившегося видео, информируя об этом генератор [9]. Позднее было создано множество различных версий архитектуры, улучшающих качество получаемых изображений (DCGAN), а также позволяющих контролировать генерацию: появилась возможность указывать класс генерируемых изображений (CGAN), стиль (StyleGAN), текстовое описание (StackGAN), подавать на вход исходное изображение, часть видео [10]. Недостатком GANs является нестабильность обучения, требующая точной настройки параметров, высокие требования к вычислительным ресурсам и большие временные затраты при обучении, а также эффект «mode collapse» – ограниченное количество сгенерированных образцов.

Примерно в то же время, что и GANs, появились Flow-based модели (например, модель GLOW). Эти модели используют серию обратимых преобразований (поток) для отображения простого базового распределения (например, гауссового шума) на сложное распределение данных [11]. Обратимость преобразований заключается в том, что можно как генерировать данные из простого распределения (например, гауссова), так и вычислять плотность данных, которые уже были сгенерированы. Основная идея — найти "поток", который соединяет шум с данными. Они быстрые и точные, а также позволяют точно оценивать правдоподобие, но требовательны к вычислительным ресурсам и обладают сложной архитектурой.

Авторегрессионные модели позволяют генерировать длинные видео благодаря последовательной генерации – каждый кадр генерируется на основе предыдущего, что позволяет увеличить согласованность кадров в продолжение видео [10]. В рамках этой концепции пространственные соотношения между фрагментами имеют решающее значение, поскольку каждый последующий фрагмент должен быть идеально выровнен с соседними, чтобы обеспечить визуальную согласованность во всем кадре. Этот метод использует пространственные зависимости, присущие видеоконтенту, гарантируя, что по мере развития видео каждый кадр остается согласованным с предыдущим не только во времени, но и в

пространстве. Однако последовательная генерация кадров связана с большими вычислительными затратами и низкой скоростью работы.

Настоящим прорывом в области генерации изображений и видео стали диффузионные модели. Идея провести аналогию между свойствами термодинамической диффузии и глубоким обучением появилась еще в 2015 году, но ее настоящее развитие началось после 2020 года. Процесс генерации в диффузионных моделях состоит из двух этапов:

- Прямой процесс: исходное изображение или видео постепенно, шаг за шагом, зашумляется добавлением гауссовского шума. После множества итераций данные превращаются в практически чистый шум;
- Обратный процесс: нейронная сеть обучается постепенно удалять шум на каждом шаге, восстанавливая данные. Именно этот процесс и используется для генерации новых видео из шума. Шум удаляется на основе как пространственного содержания отдельных кадров, так и временных соотношений между последовательными кадрами. Такой подход позволяет создавать видео, в котором каждый кадр не только визуально соответствует предыдущему, но и делает последовательность кадров плавной [9].

Из-за итеративного подхода к зашумлению и шумоподавлению процесс обучения и генерации, как правило, более стабильный, чем у GAN, восстанавливающих изображение из шума за один временной шаг, а результаты получаются качественными и разнообразными. Но это сказывается на скорости генерации.

Большинство известных современных моделей генерации видео, такие как Sora, Stable Video Diffusion, GEN-3, Kandinsky 4 и др., являются диффузионными.

Таким образом, у каждого метода есть свои достоинства и недостатки (см. табл.1), но в каждом новом методе применялись ключевые идеи предшествующих подходов, создавая тем самым основу для будущих моделей генерации видео.

Используемые методы

В данной работе исследовалась модель генерации видео Stable Video Diffusion от Stability AI [12]. SVD — это генеративная модель с открытым исходным кодом, которая специализируется на преобразовании статичных изображений в динамичные видео. Она относится к латентным диффузионным моделям [5]. Вместо работы непосредственно с пикселями кадров видео, что требует большого количества видеопамати для хранения данных при вычислениях, модель работает в сжатом латентном пространстве.

Архитектура Stable Diffusion состоит из VAE, кодирующего входное изображение в латентное пространство и восстанавливающего сгенерированное изображение до первоначального размера, и архитектуры U-Net, предсказывающая шум и отвечающая непосредственно за генерацию изображения.

Таблица 1 – Достоинства и недостатки

Table 1 – Advantages and disadvantages

Метод	Достоинства	Недостатки
VAE	Один из первых методов генерации изображений; возможность управлять генерацией.	Генерация размытых данных
GAN	Скорость генерации	Нестабильность обучения; высокие требования к вычислительным ресурсам; большие временные затраты при обучении; эффект «mode collapse»
Flow-based модели	Скорость, точность; Возможность оценки генерации.	Требовательность к вычислительным ресурсам; Сложная архитектура.
Авторегрессионные модели	Учет особенностей предыдущего кадра для генерации следующего	Большие вычислительные затраты; Низкая скорость.
Диффузионные модели	Стабильный процесс обучения и генерации; Качественный результат; Возможность управления генерацией.	Требовательность к вычислительным ресурсам; Скорость работы

При генерации видео сложность заключается в том, чтобы преобразовать шум не просто в изображение, а в серию изображений (кадров), плавно и согласованно меняющихся со временем. Работа со временем требует от модели понимания, как объекты и сцена изменяются со временем. Модель SVD инициализируется весами из уже обученной модели Stable Diffusion для изображений, но в ее архитектуре есть временные слои и слои внимания (spatial cross-attention layers) [12]. Это позволяет модели учитывать особенности других кадров для обеспечения временной согласованности. Благодаря архитектуре U-Net (skip-connections – пропускающие связи между слоями одинаковой размерности [13]) изображение рассматривается на разных уровнях детализации для обеспечения пространственной согласованности.

SVD способна генерировать видео, опираясь на исходный кадр. Модель создает параллельно 14/25 кадров (зависит от выбранной версии) благодаря временным слоям и слоям внимания. Это обеспечивает вычислительную эффективность. Однако недостатком такого подхода является передача ошибки между кадрами: ошибка в одном кадре приводит к ошибкам в других кадрах. В данной работе влияние ошибки показано на примере изменения цветовых и текстурных характеристик кадров видео.

Для уменьшения влияния ошибок на генерацию последующего сегмента видео и финальный результат было принято решение проанализировать изменение цветовых характеристик кадров по различным метрикам и выявить кадр, с которого следует начинать обработку.

Для анализа изменений цветовых и текстурных характеристик кадра использовались следующие методы:

1) Изменение цветовых координат цветовой модели HSV – Hue (тон), Saturation (насыщенность), Value (значение/яркость). Тон – это доминирующий цвет, воспринимаемый наблюдателем. Насыщенность цвета характеризует количество белого цвета в

нем; чем выше этот параметр, тем чище цвет. Яркость показывает, насколько цвет яркий или темный. Модель HSV наиболее близка к тому, как человек воспринимает цвет.

2) Сравнение гистограмм по трем каналам цветовой модели RGB. Гистограмма цифрового изображения с уровнями яркости в диапазоне $[0, L - 1]$ называется дискретная функция $h(g_k) = n_k$, где g_k есть k -й уровень яркости, а n_k — число пикселей на изображении, имеющих яркость g_k . В модели RGB каждый цвет представляется красным, зеленым и синим первичными основными цветами (компонентами). Для сравнения гистограмм использовались два метода: с помощью коэффициента корреляции и расстояния Кульбака — Лейблера. Коэффициент корреляции показывает, насколько тесно связаны распределения двух переменных: значение близкое к 1 или -1 означает сильную связь (положительную или отрицательную соответственно), а значение близкое к 0 — отсутствие связи. Он задается следующей формулой [14]:

$$d(H_1, H_2) = \frac{\sum_i (H_1(i) - \bar{H}_1)(H_2(i) - \bar{H}_2)}{\sqrt{\sum_i (H_1(i) - \bar{H}_1)^2 \sum_i (H_2(i) - \bar{H}_2)^2}},$$

где H_1, H_2 – сравниваемые гистограммы.

Расстояние Кульбака-Лейблера — это мера различия между двумя вероятностными распределениями, которая показывает, насколько одно распределение отличается от другого. Оно задается следующей формулой [14]:

$$d(H_1, H_2) = \sum_i H_1(i) \log \frac{H_1(i)}{H_2(i)},$$

где H_1, H_2 – сравниваемые гистограммы.

3) Изменение среднего значения цвета по трем каналам модели RGB [15]. Для этого находилось среднее значение цвета по изображению. Изменение среднего значения цвета находилось по формуле Евклидова расстояния между двумя точками в трехмерном пространстве:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2},$$

где $(x_1; y_1; z_1)$ и $(x_2; y_2; z_2)$ – средние значения по трем каналам для каждого изображения.

4) Расчет цветового различия ΔE в цветовой модели CIELAB (CIELCH). Система $L^*a^*b^*$ разделяет интенсивность (которая представлена яркостью L^*) и цветность (которая представлена двумя цветовыми различиями: a^* — красный минус зеленый и b^* — зеленый минус синий). В 1994 году было предложена модификация системы CIELAB – CIELCH. При этом были предложены полярные координаты цветового тона H и цветности S . Координата H обозначается как h^*_{ab} и является «угловой», а координата S обозначается как S_{ab} и является «радиальной», относительно центра координат (оси L^*). Они вычисляются по формулам [16]:

$$C^*_{ab} = \sqrt{(a^*)^2 + (b^*)^2};$$

$$h^*_{ab} = \arctg\left(\frac{b^*}{a^*}\right),$$

после чего цветовой тон переводится из радиан в градусы.

CIELCH приближает описание цвета к восприятию его человеком.

Цветовое различие ΔE рассчитывается по следующей формуле:

$$\Delta E = \left[\left(\frac{\Delta L}{K_L S_L} \right)^2 + \left(\frac{\Delta C}{K_C S_C} \right)^2 + \left(\frac{\Delta H}{K_H S_H} \right)^2 \right]^{1/2},$$

где $\Delta L = L^*_1 - L^*_2$; $\Delta C = C^*_{12} - C^*_{21}$; $\Delta H = \sqrt{\Delta a^2 + \Delta b^2 - \Delta C^2}$; $K_L = 1$, $K_C = 1$, $K_H = 1$, $S_L = 1$, $S_C = 1 + K_1 C^*_{12}$, $S_H = 1 + K_2 C^*_{12}$, $L^*_{12} = \frac{L^*_1 + L^*_2}{2}$, $C^*_{12} = \frac{C^*_1 + C^*_2}{2}$, $K_1 = 1$, $K_2 = 1$.

5) Измерение резкости изображения с помощью Лапласиана [17]. Оператор Лапласа используется в цифровой обработке изображений для обнаружения границ на изображении. Он заключается в измерении скорости изменения интенсивности пикселей на изображении: высокие значения оператора Лапласа указывают на границы между различными областями изображения. Оператор Лапласа — это оператор производной второго порядка. Он применяется к каждому пикселю изображения и рассчитывается по следующей формуле:

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}.$$

После обнаружения краев можно оценить четкость изображения. Четкое изображение будет иметь много обнаруженных краев, что приведёт к большей дисперсии значений Лапласиана по всему изображению. Чем сильнее изображение размыто, тем меньше обнаруженных краев и тем ниже значение дисперсии. Дисперсия — это мера разброса значений случайной величины относительно её математического ожидания.

6) Измерение энтропии и контраста изображения. Энтропия — это мера неопределенности или количества информации. Энтропия увеличивается, когда изображение становится более сложным, контрастным, зашумленным или детализированным. Резкость изображения также влияет на энтропию: резкий скачок значений пикселей на границе вносит высокую неопределенность в этой конкретной обла-

сти. Энтропия уменьшается, когда изображение становится более простым, однородным, размытым или гладким. Энтропия изображения находится по следующему алгоритму [18]:

- Строится гистограмма значений яркости изображения;

- Гистограмма нормализуется путем деления частоты каждого значения пикселя на общее количество пикселей, чтобы получить вероятность p_i для каждого значения яркости;

- Для расчета энтропии применяется формула энтропии Шеннона:

$$H = - \sum_i p_i \cdot \log_2 p_i;$$

Контрастность изображения — это разность между максимальным и минимальным уровнями яркости в данном изображении [19]. Контрастность и энтропия связаны: чем контрастнее изображение (резче перепады яркости), тем больше неопределенность, а соответственно и энтропия.

В данной работе для увеличения продолжительности видео генерировалось сразу несколько сегментов, опорным кадром для каждого последующего сегмента являлся последний кадр предыдущего, после чего сегменты «склеивались». Для уменьшения видимых различий между сегментами в местах склейки использовались методы интерполяции. Интерполяция — это процесс вычисления промежуточных значений на основе дискретного набора данных. В контексте видео интерполяция направлена на увеличение частоты кадров видеопоследовательности путём синтеза одного или нескольких промежуточных кадров между заданными [20]. Математически интерполяция представляет собой процесс нахождения значения функции $f(x)$ в промежуточной точке x , основываясь на известных значениях функции в соседних точках $f(x-1)$ и $f(x+1)$. Одним из ключевых параметров при интерполяции видеопотока является коэффициент α , который определяет положение промежуточного кадра относительно исходных кадров. Значение α варьируется в пределах $[0; 1]$:

$\alpha = 0$ — соответствует первому исходному кадру,

$\alpha = 1$ — соответствует второму исходному кадру,

а значения в интервале $(0; 1)$ соответствуют положению промежуточных кадров.

Для генерации нескольких кадров между двумя исходными кадрами коэффициент α задается в виде последовательности:

$$\alpha_i = \frac{i}{n+1},$$

где i — номер промежуточного кадра, n — общее количество генерируемых кадров. Таким образом, коэффициент α позволяет гибко управлять процессом интерполяции, распределяя промежуточные кадры равномерно или с учетом особенностей конкретного алгоритма.

Существует множество методов интерполяции кадров, в работе использовались следующие:

- Линейная интерполяция [21]. Для каждого пикселя получаемого в результате интерполяции кадра вычисляется взвешенная сумма соответствующих пикселей двух исходных кадров. Это описывается следующей формулой:

$g(x) = (1 - \alpha) * f_0(x) + \alpha f_1(x)$,
где α – коэффициент интерполяции, $g(x)$ – пиксель результирующего изображения, $f_0(x)$, $f_1(x)$ – пиксели исходных кадров.

- Интерполяция на основе оптического потока (одностороннего и двустороннего). Оптический поток – метод вычисления вектора перемещения объекта между двумя последовательными кадрами, которое возникает в результате движения непосредственно объектов или снимающей их камеры [20]. Таким образом, описывается движение объекта с интенсивностью $I(x, y, t)$, который по истечении времени dt перемещается на dx и dy , и его новая интенсивность становится равной $I(x+dx, y+dy, t+dt)$.

В работе использовался плотный оптический поток Гуннара Фарнебака [22]. В этом случае вычисляется смещение всех пикселей кадра относительно соседнего, минимизируя разность интенсивностей между кадрами.

Односторонний оптический поток – это подход, при котором вычисляется только прямой поток, то есть смещения от первого кадра ко второму. Алгоритм Фарнебака вычисляет вектор движения $flow$ для каждого пикселя первого кадра, создавая карты смещений для координат x и y . Эти карты используются для коррекции координат пикселей первого кадра с учетом заданного коэффициента интерполяции α :

$$\begin{cases} x_{new} = x + \alpha * flow_x \\ y_{new} = y + \alpha * flow_y \end{cases}$$

где x_{new} , y_{new} – координаты пикселей после смещения. Новый кадр создается путем «сдвига» пикселей исходного кадра на вычисленные новые координаты.

Двусторонний оптический поток помимо прямого потока вычисляет и обратный – смещения от второго кадра к первому, после чего полученные промежуточные изображения накладываются, образуя новый кадр.

Результаты экспериментов

Для генерации видео была использована модель Stable Video Diffusion (версия SVD-XT) от Stability AI. Данная модель позволяет генерировать видео с разрешением 576x1024, содержащее 25 кадров. Разработка и тестирование проводились в облачной среде Google Colab, предоставляющей GPU, на языке программирования Python. В среднем на генерацию видео, состоящего из двух сегментов по 25 кадров в каждом, общей длительностью 5 сек затрачивалось около 15 мин и 11,6 GB VRAM.

На рис. 1 представлен результат генерации видео, состоящего из 55 кадров (50 сгенерированных + 5 интерполированных), общей длительностью 5,5 сек. На вход модели генерации подавалось изображение подстилающей поверхности, полученное с дрона. Датасет был получен в рамках конкурса «Автономный поиск» [23]. Входным изображением для генерации второго сегмента видео являлся последний кадр первого сегмента. Это делалось для уменьшения влияния случайности при генерации следующего сегмента.

Несмотря на то, что место склейки двух сегментов практически незаметно человеческому глазу, было

принято решение применить интерполяцию к полученным сегментам для создания промежуточных кадров между ними.

Extracted Frames (every 4-th frame)



Рис. 1 – Сгенерированное видео

Fig. 1 – Generated video

При использовании линейной интерполяции в месте склейки наблюдается дрожание камеры и видно наложение кадров (см. рис.1, кадр 27). Это связано с математическими особенностями, лежащими в основе метода.

При интерполяции методом оптического потока наблюдается сдвиг кадра назад. Для нейтрализации данного эффекта было принято решение использовать метод двустороннего оптического потока, благодаря которому эффект сдвига кадра удалось устранить, однако в результате наложения полученных кадров возникла та же проблема, что и в случае использования линейной интерполяции.

В ходе разработки указывались следующие параметры (см. табл.2):

- версия – версия модели SVD
- torch_dtype - формат хранения и вычислений с числами с плавающей точкой – для экономии VRAM;
- variant - указание на использование предварительно сконвертированных весов – 16 бит – для экономии RAM;
- num_frames - количество генерируемых кадров;
- num_inference_steps - количество шагов шумоподавления;
- motion_bucket_id - интенсивность движения видео;
- decode_chunk_size - количество кадров, декодируемых VAE за раз;
- fps - кадровая частота при воспроизведении;
- num_segments - количество генерируемых сегментов видео;

- use_interpolation - необходимость интерполяции (True/False);
- type_interpolation - тип интерполяции ('linear', 'flow', 'biderect_flow');
- initial_image_path - путь к исходному изображению.

Таблица 2 – Настраиваемые параметры

Table 2 – Configurable parameters

Параметры модели	
версия	SVD-XT
torch_dtype	torch.float16
variant	fp16
num_frames	25
num_inference_steps	30
motion_bucket_id	80
decode_chunk_size	2
fps	10

Кадры полученного видео имеют различные цветовые и текстурные характеристики. Уже к концу первого сегмента видео (см. рис.1, кадр 20) становится заметным изменение оттенка и насыщенности, а также текстуры: контуры объектов становятся более четкими, а текстура сглаживается, что приводит к потере деталей. Чтобы не подавать на вход следующему сегменту изображение, сильно отличающееся от исходного по качеству, необходимо определить кадры, с которых нужно начать выполнять дополнительную обработку.

Для определения кадра, с которого требуется начать применять алгоритмы обработки, были использованы метрики сравнения всех кадров видео с

исходным первым кадром: изменение тона, насыщенности и яркости, сравнение гистограмм трех составляющих цвета, изменение среднего по трем каналам, расчет цветового различия ΔE , изменение четкости. Также была подсчитана энтропия и контрастность для каждого кадра.

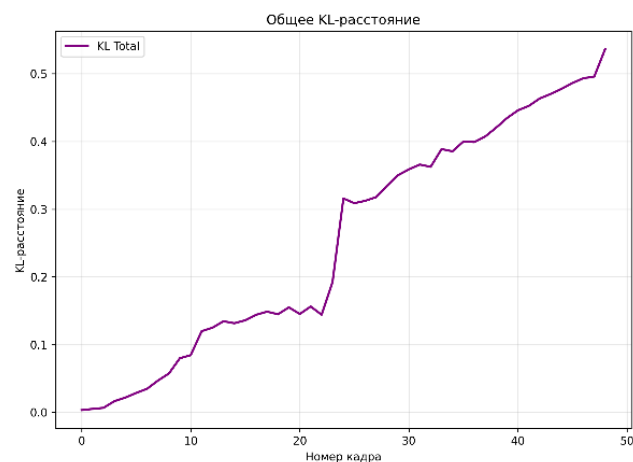


Рис. 2 – Сравнение гистограмм

Fig. 2 – Comparison of histograms

Сравнивая гистограммы по среднему расстоянию Кульбака — Лейблера (см. рис.2) можно видеть, что медленные изменения цветовых характеристик кадра происходят в первые 10 кадров, после 10 кадра наблюдается резкое изменение, а затем чего цвет стабилизируется.

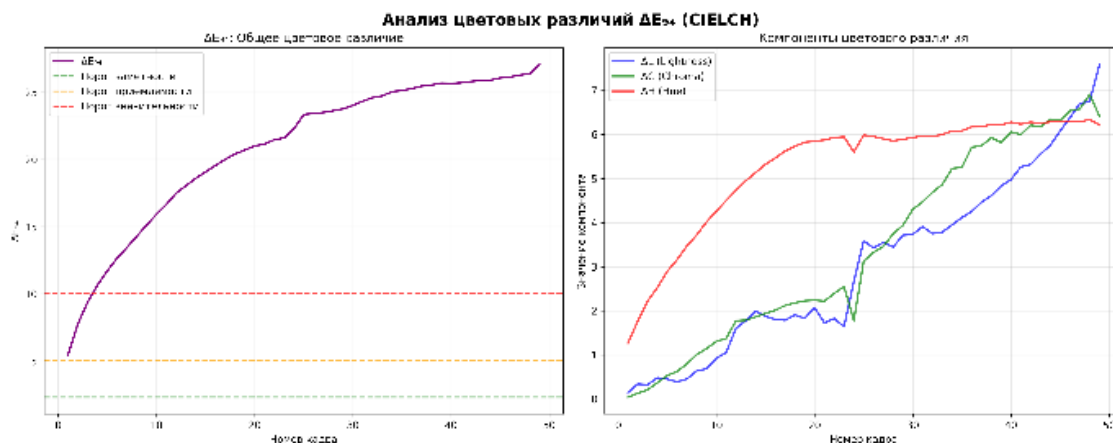


Рис. 3 – Измерение цветовых различий

Fig. 3 – Measurement of color differences

Был проведен анализ цветовых различий кадров в цветовом пространстве CIELAB. На графиках (см. рис.3) можно видеть, что ближе к 5 кадру цветовые различия становятся неприемлемыми, а такие показатели, как светлота (Lightness) и цветность (Chroma) резко меняются после 10 кадра, в то время как тон (Hue) стабильно растет.

Также было сделано измерение энтропии и контрастности изображения. На графиках (см. рис. 4) можно видеть, что энтропия растет в то время, как растет и контрастность изображения, из чего можно сделать вывод, что помимо работы, с цветовыми компонентами, также нужно провести коррекцию контрастности. Также на рис.5 показано изменение гистограммы яркости по мере увеличения контраста.

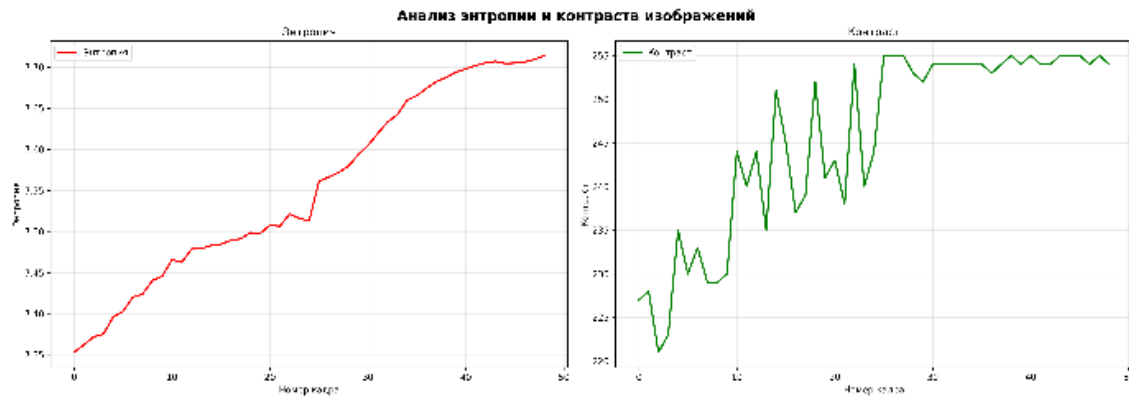


Рис. 4 – Измерение энтропии и контраста

Fig. 4 – Measurement of entropy and contrast

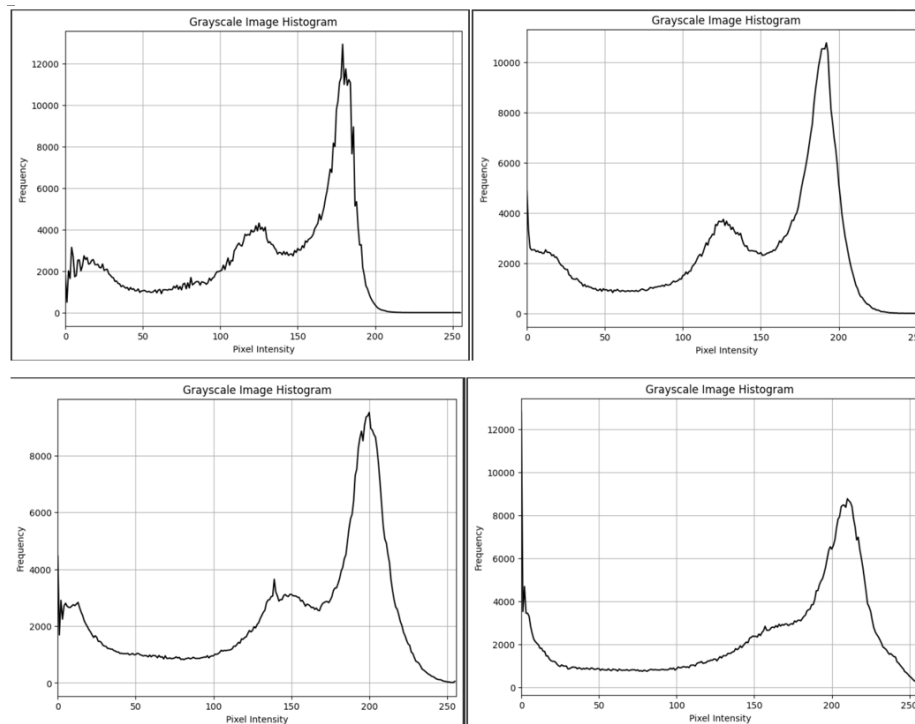


Рис. 5 – Изменение гистограммы яркости (кадр 1, 15, 30, 45)

Fig. 5 – Change in brightness histogram (frames 1, 15, 30, 45)

Анализ изменения резкости кадра с помощью оператора Лапласа показал, что четкость кадра повышается в течение видео. Это происходит из-за увеличения контраста, что делает контуры областей изображения более резкими, и приводит к увеличению энтропии.

Результаты измерения изменений компонент цветовой модели HSV и RGB так же показали, что пик изменений приходится на 10-12 кадр видео, после чего цветовые характеристики стабилизируются.

Результаты измерений, непоказанные на графиках, приведены в табл.3. В ней рассматриваются следующие показатели:

Fr – номер кадра, с которым сравнивается первый кадр (0) по различным показателям;
Hue – изменение тона (модель HSV);
Sat – изменение насыщенности (модель HSV);

Brg – изменение яркости (модель HSV);
Vcor – корреляция по каналу В (сравнение гистограмм);
Gcor – корреляция по каналу G (сравнение гистограмм);
Rcor – корреляция по каналу R (сравнение гистограмм);
Bdif – изменение среднего значения цвета по каналу В;
Gdif – изменение среднего значения цвета по каналу G;
Rdif – изменение среднего значения цвета по каналу R;
Tdif – изменение общего среднего значения цвета;
Sharp – изменение четкости (метод Лапласа);
Var – дисперсия Лапласиана (метод Лапласа).

Таблица 3 – Результаты анализа

Table 3 – Analysis results

fr	hue	sat	brg	bcor	gcor	rcor	bdif	gdif	rdif	tdif	sharp	var
1	86,30	111,10	126,56	1,00	0,99	0,99	0,48	0,32	0,38	0,68	0,00	498,18
2	106,52	121,04	131,52	0,99	0,99	0,99	1,30	0,77	0,87	1,74	-0,04	517,40
3	118,70	129,41	131,09	0,99	0,99	0,99	1,46	0,74	0,71	1,79	-0,10	546,58
4	124,07	134,39	134,61	0,96	0,97	0,98	2,27	1,19	0,98	2,74	-0,10	546,82
5	130,60	140,64	135,09	0,94	0,97	0,97	2,55	1,14	0,72	2,89	-0,20	600,08
6	132,52	142,91	134,84	0,92	0,96	0,97	2,59	0,99	0,53	2,82	-0,21	603,70
7	137,42	147,74	136,46	0,90	0,96	0,96	3,13	1,18	0,45	3,37	-0,24	619,39
8	135,49	151,47	138,76	0,87	0,93	0,95	3,92	1,72	0,52	4,31	-0,23	611,82
9	135,37	154,07	139,63	0,84	0,92	0,92	4,33	1,91	0,35	4,74	-0,28	635,84
10	134,77	156,17	142,27	0,78	0,88	0,91	5,27	2,61	0,69	5,92	-0,29	643,36
11	134,48	155,99	143,04	0,76	0,87	0,90	5,67	2,94	0,90	6,45	-0,43	714,50
12	137,51	159,22	147,06	0,65	0,81	0,88	7,81	4,42	1,74	9,14	-0,39	692,62
13	136,37	158,59	147,45	0,63	0,79	0,87	8,24	4,87	2,15	9,81	-0,61	802,21
14	141,18	157,59	147,71	0,60	0,78	0,87	9,01	5,36	2,78	10,85	-0,46	728,63
15	142,24	157,57	147,04	0,61	0,79	0,87	8,90	5,08	2,39	10,52	-0,48	738,14
16	144,29	156,87	146,92	0,60	0,79	0,87	8,83	4,89	2,27	10,35	-0,46	729,09
17	145,24	157,18	147,12	0,58	0,78	0,85	9,03	4,86	2,11	10,47	-0,49	742,51
18	145,65	157,10	147,32	0,57	0,78	0,85	9,40	5,17	2,37	10,99	-0,49	743,47
19	143,13	157,67	146,93	0,58	0,78	0,86	9,19	5,02	1,95	10,65	-0,48	736,46
20	143,36	156,65	147,53	0,55	0,77	0,85	9,80	5,66	2,56	11,60	-0,43	711,63
21	141,46	156,86	146,32	0,58	0,79	0,86	8,89	4,81	1,63	10,24	-0,45	722,58
22	147,66	156,44	147,14	0,55	0,78	0,85	9,64	5,02	1,89	11,03	-0,42	708,42
23	146,71	156,83	146,53	0,58	0,80	0,86	9,47	4,55	1,12	10,57	-0,50	747,66
24	138,45	146,87	148,51	0,51	0,67	0,74	11,05	7,20	4,18	13,83	-0,42	706,87
25	142,01	150,09	152,68	0,26	0,43	0,57	16,06	9,86	4,69	19,42	-0,44	715,72
26	150,27	150,51	151,93	0,27	0,46	0,58	16,15	9,34	4,35	19,16	-0,44	715,52
27	144,15	152,09	151,95	0,26	0,45	0,58	16,54	9,77	4,15	19,65	-0,37	683,06
28	154,76	152,43	152,17	0,24	0,46	0,57	17,04	9,42	3,89	19,86	-0,33	662,18
29	143,08	153,22	152,86	0,21	0,41	0,56	17,86	10,30	3,68	20,94	-0,36	676,05
30	153,71	152,84	152,82	0,17	0,39	0,53	18,80	10,27	3,66	21,73	-0,36	678,12
31	146,10	152,88	152,40	0,16	0,37	0,53	19,33	10,87	3,42	22,44	-0,36	676,95
32	152,21	152,42	151,53	0,14	0,36	0,52	19,52	10,43	3,04	22,34	-0,36	679,15
33	146,99	152,80	151,40	0,15	0,36	0,53	19,77	10,62	2,55	22,58	-0,39	690,04
34	157,39	152,62	151,94	0,10	0,32	0,48	21,12	10,95	3,15	24,00	-0,35	673,14
35	150,58	152,96	151,42	0,10	0,33	0,50	21,45	11,49	3,06	24,52	-0,44	717,48
36	161,08	153,71	151,05	0,06	0,31	0,48	22,78	11,77	3,23	25,84	-0,39	694,33
37	153,38	153,56	150,53	0,06	0,31	0,50	23,27	12,46	3,21	26,59	-0,51	751,31
38	161,52	153,77	150,31	0,04	0,30	0,49	24,19	12,75	3,84	27,61	-0,42	706,24
39	155,64	153,22	150,22	0,03	0,27	0,46	24,51	13,44	4,26	28,28	-0,42	707,81
40	165,51	153,24	150,70	0,01	0,26	0,43	25,51	13,69	4,89	29,36	-0,44	715,20
41	158,68	152,68	150,87	0,00	0,23	0,42	26,04	14,54	5,32	30,30	-0,46	725,68
42	165,88	152,24	150,60	-0,01	0,23	0,40	26,77	14,58	5,67	31,01	-0,44	717,70
43	156,38	152,56	150,41	-0,02	0,19	0,38	27,13	15,33	5,82	31,71	-0,45	723,00
44	164,72	152,40	150,10	-0,03	0,19	0,38	28,12	15,72	6,57	32,88	-0,45	724,75
45	155,99	152,52	149,77	-0,03	0,16	0,38	28,73	16,68	6,90	33,93	-0,43	714,00
46	165,97	152,80	149,97	-0,04	0,15	0,37	30,08	17,30	7,96	35,61	-0,47	731,31
47	159,39	153,20	150,02	-0,05	0,13	0,37	30,87	18,25	8,25	36,80	-0,43	712,56
48	173,21	153,77	149,69	-0,05	0,14	0,36	31,74	18,18	8,67	37,60	-0,44	714,90
49	155,02	151,26	149,84	-0,07	0,02	0,29	32,88	20,63	10,41	40,19	-0,49	744,71

На основе проведенного анализа был сделан вывод, что требуется проводить коррекцию как цветовых характеристик по отдельности, так и сгенерированного кадра в целом начиная с 10-13 кадра видео.

Закключение

Таким образом, генерация видео является перспективным и быстро развивающимся направлением, которое находит применение в различных областях, от креативной индустрии до автономных транспортных средств. Изучение моделей генерации видео, их возможностей и ограничений способствует как усовершенствованию уже существующих методов, так и созданию новых. В ходе исследования были изучены возможности модели генерации видео Stable Video Diffusion, а именно ее версии XT, был разработана программа генерации видео на основе исходного кадра подстилающей поверхности, проанализирован полученный результат, выявлен участок видео – 10-13 кадр, с которого следует выполнять постобработку.

В дальнейшем планируется выявление и коррекция характеристик кадров, требующих постобработки. Также будут исследованы возможности дополнительного контроля генерируемого видео с помощью текстовых запросов, а также способы контроля направления движения видео.

Исследование проведено в рамках Передовой инженерной школы «Комплексная авиационная инженерия» (Соглашение 075-15-2025-129).

Литература

1. J. Kaur. Generative Video Models | A Complete Guide (2024). [Электронный ресурс] URL: <https://www.xenonstack.com/blog/generative-video-models> (дата обращения: 15.10.2025).
2. A. Germanidis. Introducing General World Models (2023). [Электронный ресурс] URL: <https://runwayml.com/research/introducing-general-world-models> (дата обращения: 21.10.2025).
3. T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, A. Ramesh. Video generation models as world simulators (2024). [Электронный ресурс] URL: <https://openai.com/index/video-generation-models-as-world-simulators/> (дата обращения: 16.10.2025).
4. A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, G. Corrado. *arXiv:2309.17080* (2023). DOI:10.48550/arXiv.2309.17080.
5. A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. Wook Kim, S. Fidler, K. Kreis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, P. 22563-22575.
6. Bank D., Koenigstein N., Giryes R. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, Springer, 2023, P. 353-374.
7. Michelucci U. *arXiv:2201.03898* (2022). DOI:10.48550/arXiv.2201.03898.
8. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. *Advances in neural information processing systems*. 3, 11, 2672-2680 (2014). DOI:10.1145/3422622.
9. C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, L. Bai. *arXiv:2403.16407* (2024). DOI:10.48550/arXiv.2407.08428.

10. Waseem F., Shahzad M. *ACM Computing Surveys* (2024). DOI:10.48550/arXiv.2412.18688.
11. Xie Y., Cheng X. *arXiv:2502.13394* (2025). DOI:10.48550/arXiv.2502.13394.
12. A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, R. Rombach. *arXiv:2311.15127* (2023). DOI:10.48550/arXiv.2311.15127.
13. Ronneberger O., Fischer P., Brox T. В сб. *International Conference on Medical image computing and computer-assisted intervention*, Springer international publishing, Cham, 2015, P. 234-241.
14. OpenCV. Histograms. [Электронный ресурс] URL: https://docs.opencv.org/3.4/d6/dc7/group__imgproc__hist.html#ga994f53817d621e2e4228fc646342d386 (дата обращения: 17.10.2025).
15. И.А. Красная, С.Н. Михальченко, Б.С. Яковлев. *Известия Тульского государственного университета. Технические науки*, 5, 58-63 (2021). ISSN 2071-6168.
16. В. Л. Жбанова, *Светотехника*, 1, С. 36-40 (2020).
17. Bansal R., Raj G., Choudhury T. *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, 2016. P. 63-67.
18. T. Hayashi, D. Cimr, H. Fujita, R. Cimler, *Information Sciences*, 647, 3, 119539, (2023). DOI:10.1016/j.ins.2023.119539.
19. Рафаэл С. Гонсалес, Ричард Е. Вудс. *Цифровая обработка изображений, 3 изд.*, Техносфера, Москва, 2012. 1104 с.
20. D. Kye, C. Roh, S. Ko, C. Eom, J. Oh. *arXiv:2506.01061* (2025). DOI:10.48550/arXiv.2506.01061.
21. А. А. Березкин, Р. В. Киричек, З. В. Кулунчаков, А. А. Ченский. *Экономика и качество систем связи*, 2, 36, 92-107 (2025). eISSN: 2500-1833.
22. Farneback G. *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, 2003. P. 363-370.
23. Конкурс «Автономный поиск». [Электронный ресурс] URL: <https://vision.upgreat.one/> (дата обращения: 16.10.2025).

References

1. J. Kaur. Generative Video Models | A Complete Guide (2024). [Electronic resource] URL: <https://www.xenonstack.com/blog/generative-video-models> (дата обращения: 15.10.2025).
2. A. Germanidis. Introducing General World Models (2023). [Electronic resource] URL: <https://runwayml.com/research/introducing-general-world-models> (дата обращения: 21.10.2025).
3. T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, A. Ramesh. Video generation models as world simulators (2024). [Electronic resource] URL: <https://openai.com/index/video-generation-models-as-world-simulators/> (дата обращения: 16.10.2025).
4. A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, G. Corrado. *arXiv:2309.17080* (2023). DOI:10.48550/arXiv.2309.17080.
5. A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. Wook Kim, S. Fidler, K. Kreis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, P. 22563-22575.
6. Bank D., Koenigstein N., Giryes R. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, Springer, 2023, P. 353-374.
7. Michelucci U. *arXiv:2201.03898* (2022). DOI:10.48550/arXiv.2201.03898.
8. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. *Advances*

- in neural information processing systems. 3, 11, 2672-2680 (2014). DOI:10.1145/3422622.
9. C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, L. Bai. *arXiv:2403.16407* (2024). DOI:10.48550/arXiv.2407.08428.
 10. Waseem F., Shahzad M. *ACM Computing Surveys* (2024). DOI:10.48550/arXiv.2412.18688.
 11. Xie Y., Cheng X. *arXiv:2502.13394* (2025). DOI:10.48550/arXiv.2502.13394.
 12. A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, R. Rombach. *arXiv:2311.15127* (2023). DOI:10.48550/arXiv.2311.15127.
 13. Ronneberger O., Fischer P., Brox T. В сб. *International Conference on Medical image computing and computer-assisted intervention*, Springer international publishing, Cham, 2015, P. 234-241.
 14. OpenCV. Histograms. [Электронный ресурс] URL: https://docs.opencv.org/3.4/d6/dc7/group__imgproc__hist.html#ga994f53817d621e2e4228fc646342d386 (дата обращения: 17.10.2025).
 15. I.A. Krasnaya, S.N. Mikhalechenko, B.S. Yakovlev. *News of Tula State University. Technical Sciences*, 5, 58-63 (2021). ISSN 2071-6168.
 16. V. L. Zhanova, *Lighting Technology*, 1, pp. 36-40 (2020).
 17. Bansal R., Raj G., Choudhury T. *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, 2016. P. 63-67.
 18. T. Hayashi, D. Cimr, H. Fujita, R. Cimler, *Information Sciences*, 647, 3, 119539, (2023). DOI:10.1016/j.ins.2023.119539.
 19. Rafael S. Gonzalez, Richard E. Woods. *Digital Image Processing*, 3rd edition, Technosphere, Moscow, 2012. 1104 pp.
 20. D. Kye, C. Roh, S. Ko, C. Eom, J. Oh. *arXiv:2506.01061* (2025). DOI:10.48550/arXiv.2506.01061.
 21. A. A. Berezkin, R. V. Kirichek, Z. V. Kulunchakov, A. A. Chensky. *Economics and Quality of Communication Systems*, 2, 36, 92-107 (2025). eISSN: 2500-1833.
 22. Farneback G. *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, 2003. P. 363-370.
 23. Autonomous Search Competition. [Electronic resource] URL: <https://vision.upgreat.one/> (accessed: 16.10.2025).

© Д. А. Колобова – студент кафедры Автоматизированных систем обработки информации и управления (АСОИУ), Казанский национальный исследовательский технический университета им. А.Н. Туполева (КНИТУ им. А.Н. Туполева), Казань, Россия, darya.kolobova@inbox.ru; М. П. Шлеймович – к.т.н., доцент, заведующий кафедрой АСОИУ, КНИТУ им. А.Н. Туполева, shlch@mail.ru.

© D. A. Kolobova – Student of the Department of Automated Information Processing and Management Systems (AIPMS), Kazan National Research Technical University named after A.N. Tupolev (KNRTU named after A.N. Tupolev), Kazan, Russia, darya.kolobova@inbox.ru; M. P. Shleimovich – PhD (Technical Sci.), Associate Professor, Head of the AIPMS department, KNRTU named after A.N. Tupolev, shlch@mail.ru.

Дата поступления рукописи в редакцию – 05.11.25.

Дата принятия рукописи в печать – 23.11.25