ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.89

ПРИМЕНЕНИЕ МНОГОЭТАПНОЙ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ ПО ТЕКСТОВОМУ ОПИСАНИЮ

Дроздов Дмитрий Сергеевич
Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
dmtr636@gmail.com

Федоров Виктор Олегович
Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
fedorov_vo@bmstu.ru

Аннотация

В данной статье описывается архитектура многоэтапной генеративно состязательной сети, которая используются для генерации изображений по текстовому описанию. Данная модель состоит из многоэтапной композиции блоков, состоящих из генератора и дискриминатора. Первый генератор создает грубую форму изображения на основе текстовых описаний, а второй генератор уточняет детали и создает финальное изображение. Дискриминатор оценивает качество изображений, и обучение происходит путем улучшения качества изображений для их более точной классификации дискриминатором. В процессе обучения генератор старается создать изображение, которое будет максимально близко к своему текстовому описанию, тогда как дискриминатор старается различать настоящие изображения от сгенерированных. Многоэтапная архитектура стековых генеративно состязательных сетей позволяет избежать проблем с градиентным затуханием, что обеспечивает более стабильное обучение и повышение качества сгенерированных изображений.

Ключевые слова: генеративно-состязательные сети, генерация изображений, глубокое обучение

APPLICATION OF STACKED GENERATIVE ADVERSARIAL NEURAL NETWORK FOR GENERATION OF IMAGES FROM A TEXT DESCRIPTION

Drozdov Dmitry Sergeevich
Federal State Budgetary Educational Institution of Higher Education «Bauman Moscow State Technical University» (Kaluga Branch)
dmtr636@gmail.com

Fedorov Victor Olegovich
Federal State Budgetary Educational Institution of Higher Education «Bauman Moscow State Technical University» (Kaluga Branch)
fedorov_vo@bmstu.ru

ABSTRACT

This article describes the architecture of a multi-stage generative adversarial network, which is used to generate images from a text description. This model consists of a multi-stage composition of blocks consisting of a generator and a discriminator. The first generator creates a rough shape of the image based on text descriptions, while the second generator refines the details and creates the final image. The discriminator evaluates the quality of the images, and learning occurs by improving the quality of the images for their more accurate classification by the discriminator. During the learning process, the generator tries to create an image that will be as close as possible to its textual description, while the discriminator tries to distinguish between real images and generated ones. The multi-stage architecture of stacked generative adversarial networks avoids problems with gradient decay, which provides more stable training and improved quality of generated images.

Keywords: generative adversarial networks, image generation, deep learning

В мире машинного обучения и искусственного интеллекта активно исследуются новые методы генерации изображений. Одним из наиболее перспективных подходов является использование многоэтапных генеративно-состязательных сетей, которые позволяют создавать новые изображения или другие данные на основе существующих [1]. Особый интерес вызывает возможность генерации изображений по текстовому описанию. Возможности многоэтапных генеративно-состязательных сетей могут оказаться полезными в разработке новых алгоритмов машинного зрения и распознавания образов. В данной статье будут рассмотрены принципы работы многоэтапных генеративно-состязательных сетей и их применение в генерации изображений по текстовому описанию.

Целью исследования является разработка архитектуры многоэтапной генеративно-состязательной нейронной сети, используемой для генерации изображений по текстовому описанию.

Архитектура модели относится к классу генеративно-состязательных нейронных сетей (GAN). На рисунке 1 изображена базовая архитектура генеративно-состязательной сети, используемой для генерации изображений по текстовому описанию [2].

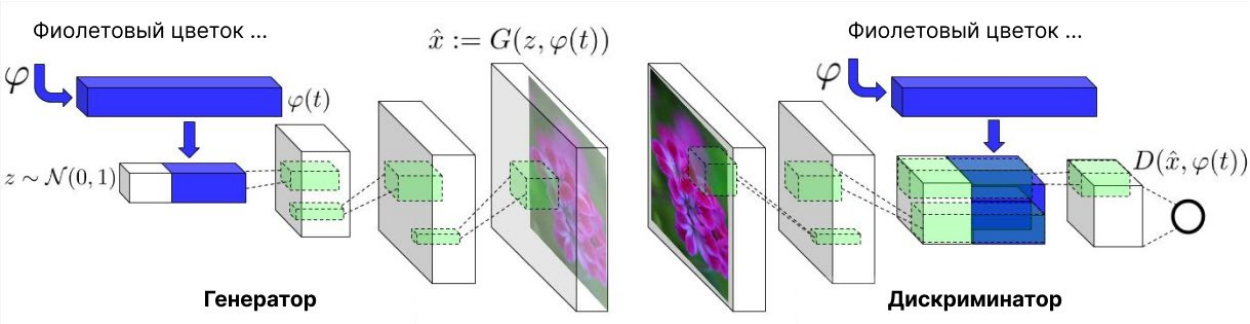

Рисунок 1. Генерация изображений по текстовому описанию с помощью GAN

Основная идея GAN заключается в том, что она состоит из двух нейронных сетей: генеративной и дискриминаторной. Генеративная сеть используется для создания новых данных на основе шумового входа, а дискриминаторная сеть используется для оценки созданных данных и сравнения их с реальными данными из обучающего набора.

В данной работе рассматривается усовершенствованный вариант многоэтапной генеративно-состязательной сети. Представленная модель состоит из двух генераторов и двух дискриминаторов. Первый генератор занимается генерацией более низкоразрешенного изображения (64x64 пикселя), а второй - более высокоразрешенного изображения (256x256 пикселей) [3].

Первый генератор выстраивается на основе DCGAN (Deep Convolutional Generative Adversarial Networks), который состоит из нескольких слоев свертки и слоев пакетной нормализации (рис. 2). Он генерирует изображения разрешением 64x64, которые затем поступают на вход первого дискриминатора.

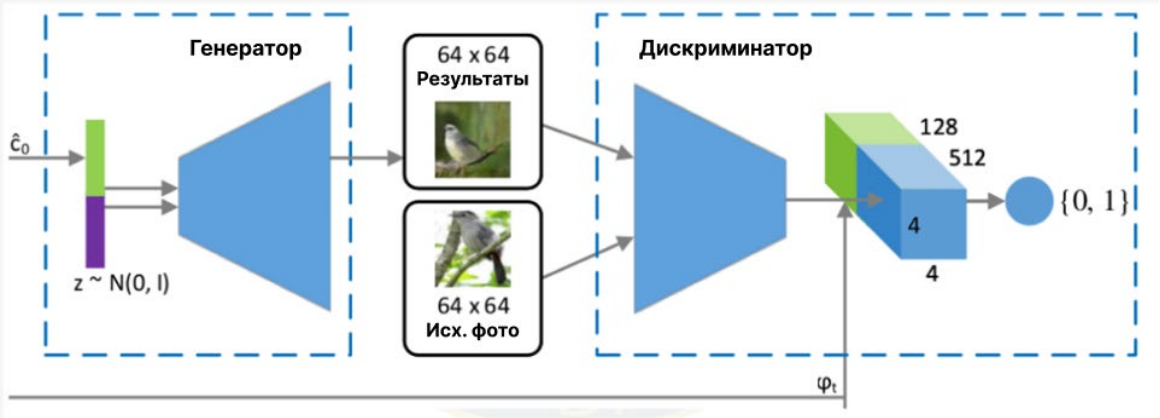

Рисунок 2. Генератор и дискриминатор 1-го этапа

Для создания таких изображений модель использует условие, которое задается вектором скрытых параметров z и картинкой-условием (текстовое описание, тег, класс объекта). Для этого в модели предусмотрена патч-дискриминаторная сеть, которая отвечает за оценку правдоподобия сгенерированных изображений [4]. Если изображение недостаточно качественное, то модель исправляет ошибки и работает до тех пор, пока не получит требуемый результат.

Второй генератор использует первый генератор и генерирует высокоразрешенное изображение (рис. 3). Он также основан на DCGAN, и состоит из множества слоев, включая нормализацию по батчам, слои свертки, слои повышения разрешения (upsampling layers) и слои активации (рис. 4) [5]. Второй генератор получает изображение разрешением 64x64 в качестве входа и выдает изображение разрешением 256x256.

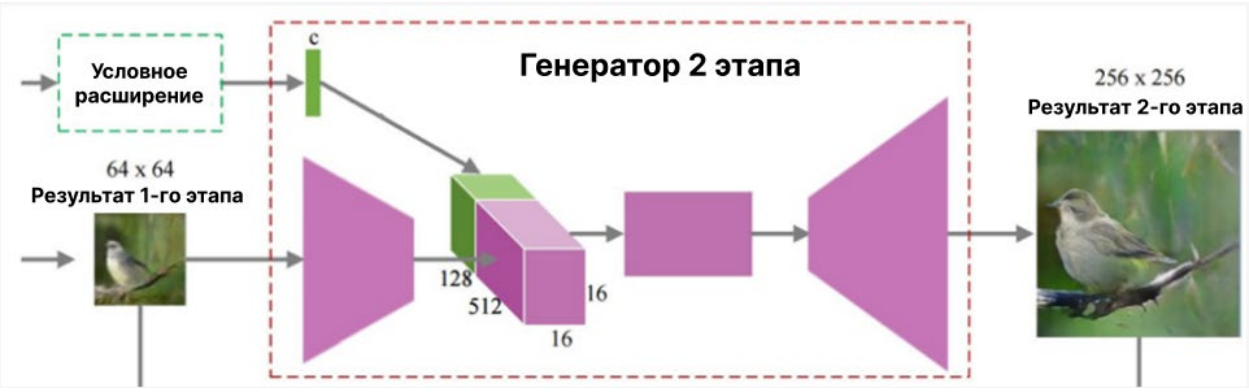

Рисунок 3. Генератор 2-го этапа

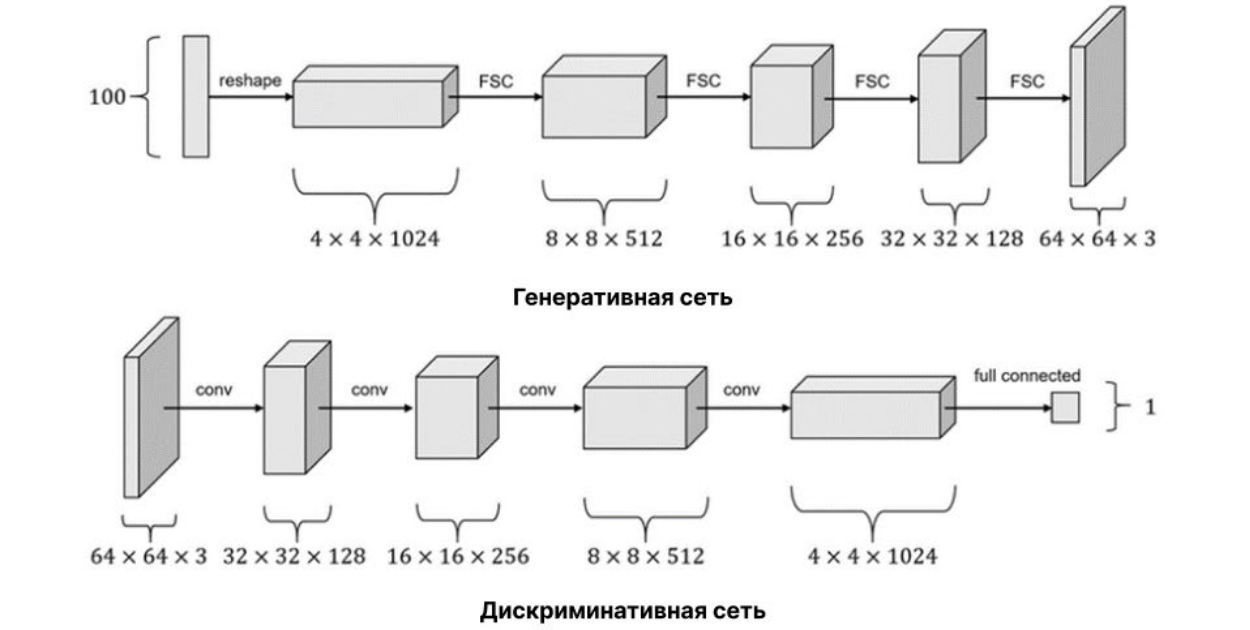

Рисунок 4. Архитектура DCGAN

Второй генератор в модели StackGAN является условной генеративной моделью, которая генерирует фотореалистичные изображения высокого разрешения по условию, заданному на первом уровне генератором. Основная задача второго генератора - улучшение качества изображений, полученных на первом уровне генератора.

Алгоритм работы второго генератора в модели StackGAN можно описать следующим образом:

1. Входными данными для второго генератора является вектор скрытых переменных, полученных на первом уровне генератора, и условие, заданное на первом уровне, которое описывает фотографию искомого объекта.

2. Сначала проводится через низкого разрешения (LR) - небольшое изображение, полученное на первом уровне генератора - через сверточную нейронную сеть, чтобы получить высокое разрешение (HR). Это позволяет сохранить и обрабатывать более точную информацию о текстуре и структуре изображения на более высоком разрешении.

3. Далее второй генератор производят обратное преобразование вектора высокого разрешения (HR) обратно в изображение. В течение этого процесса, внутри генератора строятся блоки Residual Block и Up-Sampling Blocks, которые предназначены для повышения качества изображения и сохранения ограничений на полученные изображения, заданные на первом уровне.

4. Наконец, происходит вывод изображения, которое представляет собой фотореалистичное изображение, соответствующее заданному условию.

В целом, второй генератор модели StackGAN имеет задачу сгенерировать изображения высокого качества с помощью использования сверточных нейронных сетей, которые позволяют автоматизировать и оптимизировать этот процесс. Кроме того, благодаря условным ограничениям, заданным на первом уровне генератора, изображения, полученные на втором уровне, соответствуют условиям, заданным на первом уровне, что обеспечивает контроль над процессом генерации изображений.

Дискриминатор — это сверточная нейронная сеть, которая определяет, является ли изображение реалистичным. Дискриминатор получает на вход изображение и текстовое описание, а затем проходит через несколько сверточных слоев с активационными функциями LeakyReLU и один полносвязный слой [6]. Окончательный результат выводится

в виде скалярной величины, которая определяет вероятность того, что изображение является реалистичным.

Дискриминатор обучается на реальных и сгенерированных изображениях. Он выступает в роли классификатора, который различает настоящие изображения от сгенерированных. Дискриминатор представлен в виде сверточной нейронной сети, состоящей из нескольких слоев свертки и субдискретизации.

Общая идея модели заключается в том, что первый генератор генерирует изображение с низким разрешением, которое представляет общую структуру и форму изображения. Второй генератор работает с этим изображением, чтобы улучшить его разрешение и детали. Дискриминатор оценивает, насколько хорошо сгенерированные изображения соответствуют реальным изображениям и отправляет обратную связь генераторам, чтобы они могли улучшить свою работу.

Представленную модель можно улучшить, объединив последовательно несколько генераторов и дискриминаторов [7]. На рисунке 5 представлена архитектура модели, состоящей из 3-х этапов, синтезирующих изображения размерами 64x64, 128x128, 256x256 соответственно. Сгенерированные изображения на каждом этапе поступают на вход высокоразрешающего дискриминатора JCU, который оценивает качество генерации изображения.

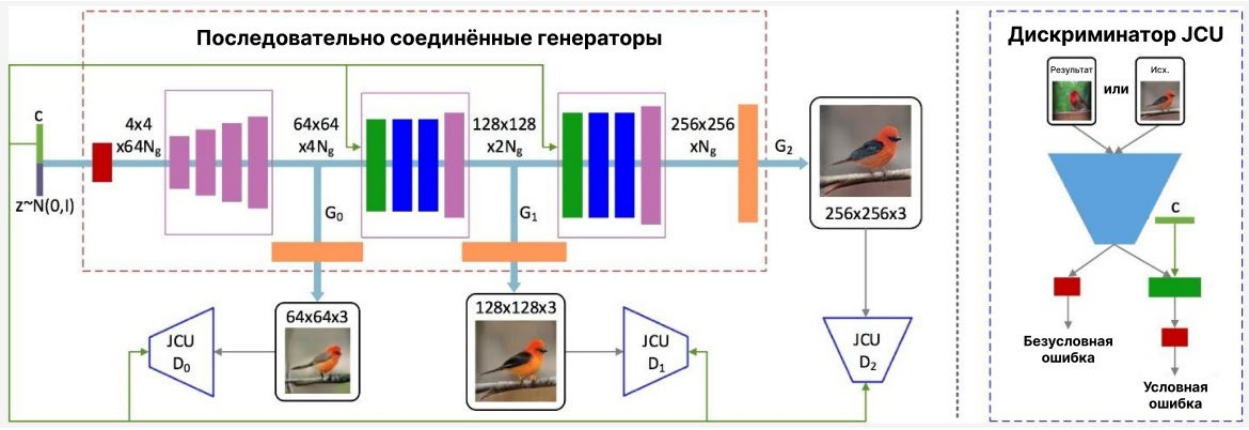

Рисунок 5. Архитектура модели с 3-мя последовательными этапами

Алгоритм обучения модели можно разделить на две части [8]:

1. Обучение первого этапа модели (Stage I):

Шаг 1. Исходными данными для этапа Stage I являются наборы разреженных сверточных карт (выходы сверточных слоев) и произвольного вектора шума, который был создан случайным образом (или вручную заранее определен).

Шаг 2. Данная модель состоит из генератора и дискриминатора. Генератор отвечает за генерацию изображений из шума, а дискриминатор отвечает за определение, насколько реалистичны сгенерированные изображения.

Шаг 3. Вначале обучается дискриминатор на реальных изображениях из тренировочного набора данных. Затем в этот дискриминатор подаются сгенерированные изображения, и он должен определить, являются ли они настоящими или нет.

Шаг 4. Далее, генератор принимает входные данные из этапа шума и старается сгенерировать изображения, которые не могут быть отличены от настоящих изображений. Генератор оптимизирует свои параметры, чтобы улучшить качество сгенерированных изображений.

2. Обучение второго этапа модели (Stage II):

Шаг 1. Генератор из первого этапа модели загружается в качестве начального приближения.

Шаг 2. Далее, мы генерируем новый набор векторов шума, который используется специальным образом в структуре модуля на этом этапе.

Шаг 3. Сгенерированные изображения с первого этапа включаются во входной набор данных для генератора второго этапа. Теперь генератор второго этапа должен сгенерировать изображения более высокого разрешения, используя в качестве контекста изображения, сгенерированные в первом этапе.

Шаг 4. Второй этап состоит из генератора, который создает изображение высокого разрешения на основе входных данных, и дискриминатора, который определяет, насколько сгенерированные изображения реалистичны.

Шаг 5. Во время обучения генератора и дискриминатора много раз оптимизируются, чтобы достичь наилучших результатов.

Шаг 6. Наконец, после окончания обучения, модель может быть использована для генерации новых, реалистичных изображений на основе входных данных из шума.

В конце процесса обучения модели необходимо совершить тестирование и оптимизацию. Этот шаг включает оценку качества изображений, созданных моделью, и оптимизацию параметров модели для улучшения ее производительности.

Количество эпох, необходимых для того, чтобы обучить модель может варьироваться от нескольких десятков до сотен в зависимости от набора данных, выбранного для обучения. В то же время, количество эпох также зависит от используемого оборудования и оптимизатора [9].

Обычно для обучения модели требуется от 50 до 100 эпох на наборах данных, таких как CIFAR-10, Fashion-MNIST или MNIST [10]. Но для более сложных наборов данных, таких как ImageNet, MS COCO, необходимо обычно более 100 эпох.

Однако, для достижения наилучшей производительности модели и создания высококачественных изображений, часто используются несколько методов, например, обучение с учителем и обучение без учителя [11]. Эти методы позволяют минимизировать ошибку в процессе обучения и сократить необходимое количество эпох для достижения необходимых результатов.

В данной статье была представлена архитектура многоэтапной генеративно-состязательной нейронной сети, используемой для генерации изображений по текстовому описанию. Описан процесс обучения представленной модели. Представленная модель обеспечивает значительное улучшение качества изображений по сравнению с предыдущими моделями.

Список литературы:

1. Дроздов Д.С., Белов Ю.С. Обзор подходов к построению моделей для генерации изображений по текстовому описанию // Фундаментальные и прикладные исследования. Актуальные проблемы и достижения: сборник статей всероссийской научной конференции, Пермь, 10 февраля 2023 г. – С. 28-30.
2. Айрапетов А.Э., Коваленко А.А. Виды генеративно-состязательных сетей // Достижения науки и образования, 2019. – №4 (45). – С. 8-13.
3. Berrahal M., Azizi M. Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques // Indones. J. Electr. Eng. Comput. Sci. 2022, pp. 972–979.
4. Zhang Y., Han S., Zhang Z. CF-GAN: Cross-domain feature fusion generative adversarial network for text-to-image synthesis // Vis. Comput. 2022, pp. 1–11.

5. Cai Y., Wang X., Yu Z. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network // IEEE Access 2019, pp. 183706–183716.
6. Lin C.H., Yumer E., Wang O. ST-GAN: Spatial transformer generative adversarial networks for image compositing // Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9455–9464.
7. Liu, R., Ge Y., Choi C.L. Diverse conditional image synthesis via contrastive generative adversarial network // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021, pp. 16377–16386.
8. Ming D., Zhuoyi Y., Wenyi H. Mastering text-to-image generation via transformers // Advances in Neural Information Processing Systems. 2021, Vol. 34, pp. 19822-19835.
9. El-Nouby A., Sharma S., Schulz H. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction // IEEE/CVF International Conference on Computer Vision (ICCV). 2019, pp. 10303-10311.
10. Zhang H., Xu T., Li H. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks // Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5907–5915.
11. Roth K., Lucchi A., Nowozin S. Stabilizing training of generative adversarial networks through regularization // Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 2018–2028.

References:

1. Drozdov D.S., Belov Yu.S. Review of approaches to building models for generating images from text descriptions // Fundamental and Applied Research. Current problems and achievements: collection of articles from the All-Russian scientific conference, Perm, February 10, 2023 – pp. 28-30.
2. Airapetov A.E., Kovalenko A.A. Types of generative adversarial networks // Achievements of science and education, 2019. – No. 4 (45). – P. 8-13.
3. Berrahal M., Azizi M. Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques // Indones. J. Electr. Eng. Comput. Sci. 2022, pp. 972–979.
4. Zhang Y., Han S., Zhang Z. CF-GAN: Cross-domain feature fusion generative adversarial network for text-to-image synthesis // Vis. Comput. 2022, pp. 1–11.
5. Cai Y., Wang X., Yu Z. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network // IEEE Access 2019, pp. 183706–183716.
6. Lin C.H., Yumer E., Wang O. ST-GAN: Spatial transformer generative adversarial networks for image compositing // Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9455–9464.
7. Liu, R., Ge Y., Choi C.L. Diverse conditional image synthesis via contrastive generative adversarial network // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021, pp. 16377–16386.
8. Ming D., Zhuoyi Y., Wenyi H. Mastering text-to-image generation via transformers // Advances in Neural Information Processing Systems. 2021, Vol. 34, pp. 19822-19835.

9. El-Nouby A., Sharma S., Schulz H. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction // IEEE/CVF International Conference on Computer Vision (ICCV). 2019, pp. 10303-10311.
10. Zhang H., Xu T., Li H. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks // Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5907–5915.
11. Roth K., Lucchi A., Nowozin S. Stabilizing training of generative adversarial networks through regularization // Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 2018–2028