УДК 378:004
DOI: 10.25686/978-5-8158-2474-4-2025-828-834

Влияние объема синтетически сгенерированных изображений в наборе данных для задачи сегментации аксонов в электронной микроскопии мозга

А. И. Миронов, Н. А. Соколов, А. А. Серебрякова
Исследовательский центр в области искусственного интеллекта, Нижегородский государственный университет им. Н. И. Лобачевского, Нижний Новгород, Россия

Аннотация. В данной работе исследуется влияние объема синтетических данных на качество сегментации изображений электронной микроскопии мозга. В качестве исходного набора данных использованы широко применяемый набор EPFL и разметка ITMM на шесть классов. Для генерации синтетических данных использовались геометрические алгоритмы, а для решения задачи сегментации применена модификация модели U-Net. Основное внимание уделено анализу зависимости точности сегментации аксонов и постсинаптических уплотнений от соотношения реальных и синтетических данных. Применение синтетических данных продемонстрировало значительное улучшение качества сегментации при ограниченном объеме реальных данных.

***Ключевые слова*:** синтетические данные, медицинские данные, сегментация, электронная микроскопия

The effect of the volume of synthetically generated images in the dataset for the task of axon segmentation in brain electron microscopy

A. I. Mironov, N. A. Sokolov, A. A. Serebryakova
Research Center for Artificial Intelligence, Lobachevsky State University of Nizhny Novgorod, Russia

Abstract. This paper examines the impact of synthetic data volume on the quality of brain electron microscopy image segmentation. The widely used EPFL dataset and the six-class ITMM labeling were used as the initial dataset. Geometric algorithms were used to generate the synthetic data, and a modification of the U-Net model was applied to solve the segmentation problem. The primary focus is on analyzing the dependence of axon and postsynaptic densification accuracy on the ratio of real to synthetic data. The use of synthetic data demonstrated a significant improvement in segmentation quality with a limited real data volume.

***Keywords*:** Synthetic data, Medical data, Segmentation, Electron microscopy, U-Net, Axon

Введение

Современные методы компьютерного зрения, особенно в области сегментации изображений, значительно продвинулись благодаря использованию глубоких нейронных сетей. Семантическая сегментация — это задача компьютерного зрения, в которой каждому пикселю изображения присваивается метка, соответствующая классу объекта, что позволяет выделить и идентифицировать различные объекты и области на изображении. Однако успешное применение моделей часто ограничивается доступностью размеченных данных, особенно в специализированных областях, таких как медицина.

В последние годы синтетические данные стали важным инструментом для преодоления этих ограничений. Их генерация позволяет создавать разнообразные и реалистичные примеры, которые могут использоваться для обучения нейронных сетей [1, 3]. Однако влияние объема синтетических данных на качество сегментации остается недостаточно изученным. Недостаточное количество данных может привести к низкой точности сегментации, в то время как избыток некачественной синтетики приведет к снижению качества сегментации на реальных данных. В контексте сложных медицинских изображений это особенно актуально, так как нет генераторов синтетики, точно воспроизводящей настоящие данные. Также в медицинских изображениях, как правило, существует сильный дисбаланс классов [4], что может приводить к смещению прогнозируемых вероятностей в сторону доминирующих классов и низкой точности распознавания редких объектов. Улучшение распознавания редких объектов может быть достигнуто при помощи добавления синтетических изображений в обучающий набор данных.

Данная работа фокусируется на применении синтетических данных для сегментации изображений электронной микроскопии мозга. Особое внимание уделяется анализу зависимости качества

сегментации аксонов и постсинаптических уплотнений от соотношения реальных и синтетических данных. Аксоны в датасете EPFL 5 являются самым редким классом, из-за чего их классификация становится невозможной без синтетических данных.

Мы ставим эксперименты, чтобы определить оптимальный баланс, обеспечивающий наибольшую точность сегментации при добавлении минимальных объемов синтетических данных.

Постановка задачи

1. Датасет

В исследовании используется Electron Microscopy Dataset, предоставленный исследовательским университетом Швейцарии EPFL 5 вместе с разметкой ITMM 6. Датасет представляет участок размером 5×5×5 микрометров, взятый из области гиппокампа CA1 мозга, что соответствует объёму 1065×2048×1536. Все три доступные разметки датасета работают с двумя выделенными частями разрешением 1024x768x165. Мы используем разметку ITMM 6, в которой, несмотря на малое количество размеченных слоев (70 из 330), они размечены на шесть классов, в отличие от двух других доступных разметок для одного класса митохондрий.

Для обучения модели было выбрано 10 слоёв, разбитых на тайлы размером 256×256. Каждый слой разбивался при помощи окна 256×256 и смещения 128 пикселей. На изображениях с аннотациями определенного класса наличие объекта отмечается белым цветом, а отсутствие – черным. Пример тайла представлен на рисунке 1.

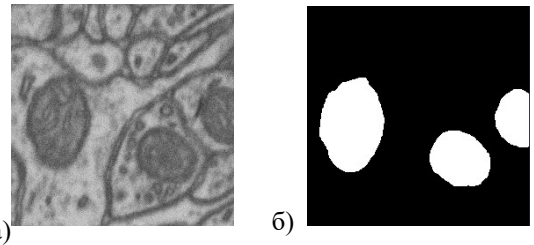


Рисунок 1. Пример тайла данных из датасета "EPFL Electron Microscopy Dataset": а – входной тайл 256×256; б – размеченная маска митохондрий

Для тестирования использовались 5 слоёв, которые в процессе тестирования также разбиваются на тайлы со смещением, а потом предсказание собирается в бинарную маску размера исходного слоя.

2. Синтетические данные

Для генерации синтетических данных электронной микроскопии использован параметризуемый алгоритм, основанный на геометрических правилах. Формы органелл моделируются с использованием случайной выборки параметров, таких как размер, положение и ориентация, что позволяет эффективно создавать разнообразные и в то же время контролируемые изображения.

Процесс генерации включает четыре этапа:

1) синтез органелл (аксоны, митохондрии, PSD (postsynaptic density, постсинаптическое уплотнение), везикулы);
2) размещение объектов на изображении с учётом отсутствия пересечений;
3) генерация мембран;
4) добавления размытия и шума.

Синтетические данные генерировались с конфигурациями, включающими следующее количество объектов: митохондрии — 0, а количество аксонов, PSD и областей везикул может варьироваться до 3 на одном изображении. Эти объекты были случайным образом распределены на изображениях размером 256×256. Более подробное описание алгоритма описано в статье [7]. Генератор изображений доступен в репозитории GitHub 8. Преимуществом такого подхода является наличие точной разметки сразу для всех классов в виде масок, а также эффективность генерации с минимальными вычислительными затратами, по сравнению, например, с диффузионными нейросетями. Генерация одного геометрического тайла занимает примерно 3 секунды на центральном процессоре i7-13700k

3400MHz, тогда как генерация тайла с использованием диффузионной модели занимает 15 секунд с использованием NVIDIA A100. На рисунке 2 приведен пример сгенерированного изображения.

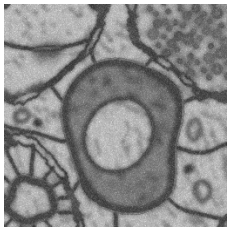
а)  б) 

Рисунок 2. Пример синтетических данных: а – сгенерированное изображение; б – сгенерированная маска

3. Tiny U-Net

U-Net 9 — это архитектура нейронной сети, изначально предложенная для задач сегментации медицинских изображений, в частности полученных с помощью микроскопии. Разработанная в 2015 году для сегментации биомедицинских изображений архитектура зарекомендовала себя [10] как одна из самых эффективных для обработки изображений с небольшим количеством размеченных данных.

В работе используется модификация классической архитектуры — Tiny U-Net [11]. Модификация имеет меньшее количество параметров, что уменьшает вычислительные затраты как на этапе обучения, так и на этапе инференса. Кроме того, Tiny U-Net лучше подходит для задач, где количество обучающих данных сильно ограничено. Используемая реализация нейронной сети находится в репозитории GitHub [12].

4. Оценка качества

Для оценки качества сегментации использовался коэффициент Dice–Score (DSC, Dice), широко применяемый в медицинской сегментации изображений. Значения метрики варьируются от 0 до 1. Пусть TP — количество пикселей, правильно классифицированных как принадлежащие целевому классу (True Positive); TN — правильно классифицированные фоны (True Negative); FP — ошибочно отнесённые к целевому классу (False Positive); FN — ошибочно отнесённые к фону (False Negative). Тогда метрика Dice определяется следующим образом:

$$DSC = \frac{2TP}{2TP + FP + FN}. \quad (1)$$

Так как в работе рассматривается многоклассовая сегментация, применяется векторная форма метрики Dice — отдельное значение DSC_i для каждого класса. Чтобы использовать её как функцию потерь при обучении нейронной сети, необходимо агрегировать её в скалярную величину. Для этого применяется линейная комбинация

$$Loss = \sum_{i=1}^{N} \alpha_i (1 - DSC_i), \alpha_i >= 0, \sum_{i=1}^{N} \alpha_i = 1, \quad (2)$$

где *Loss* — итоговая скалярная функция потерь; *N* — количество классов; α_i — весовой коэффициент для *i*-го класса, выбранный равным $\frac{1}{N}$.

Эксперименты

Для определения оптимального количества синтетических изображений, необходимого для достижения высокого качества сегментации, были выполнены следующие шаги.

1. *Подготовка данных*. Для обучения модели было выбрано 10 слоёв. Каждый слой разбили на 1170 тайлов размерностями 256x256 пикселей. Разбиение проводилось для каждого слоя с шагом 128 пикселей, в результате итоговые изображения перекрываются. Входные данные содержат вручную размеченные маски для каждого класса.

2. *Генерация синтетических данных*. С использованием геометрического генератора создано 300 синтетических изображений размером 256×256.

3. *Формирование обучающих данных*. Обучающие наборы формировались с числом настоящих данных 0, 585 и 1170 тайлов (соответствует 0, 5 и 10 изображениям). В каждый настоящий набор добавлялись синтетические изображения в объёме от 0 до 300 тайлов с шагом 30.

4. *Подсчёт органелл*. Для каждого тайла определялись количество органелл и их суммарная площадь как в синтетических, так и в реальных данных. Количество пикселей конкретного объекта считалось по маскам.

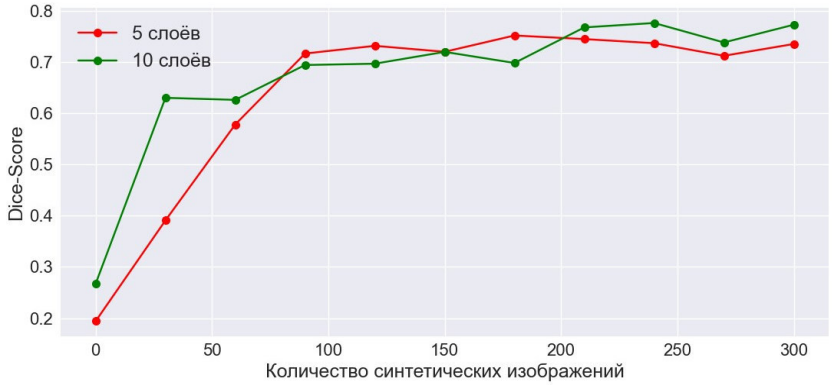
5. *Обучение модели*. Модель Tiny U-Net обучалась на каждом из наборов 5 раз с различным значением генератора случайных чисел (seed) для устойчивости оценки.

6. *Тестирование модели*. После обучения каждая модель тестировалась на выделенной тестовой выборке, качество сегментации оценивалось с помощью Dice-Score. Метрики усреднялись.

Проведено два эксперимента. Они различаются в способе генерации синтетических данных. В первом эксперименте синтетические данные содержали только аксоны, во втором – аксоны, постсинаптические уплотнения и везикулы. Выбор именно таких классов для создания геометрической синтетики обусловлен стартовыми условиями исходного набора данных EPFL и особенностями расположения компартментов относительно друг друга. В тренировочном наборе EPFL аксон представлен на 36 слоях и совершенно не похож на аксон в тестовом наборе данных. В этом случае мы ограничены в использовании других методов расширения датасета, таких как аугментация и обучаемые нейросетевые модели, в том числе диффузионные [13]. В этом же наборе данных интенсивность PSD пересекается с интенсивностью миелиновой оболочки аксона, поэтому нейросети сложно различать эти два класса. Класс везикулы появляется в синтетических данных благодаря тому, что в реальных биологических структурах скопления везикул располагаются рядом с PSD.

Результаты экспериментов

Результаты первого эксперимента на рисунках 3 и 4 показывают положительную зависимость качества сегментации от количества добавленных синтетических изображений. Красный цвет отображает эксперимент, в котором использовалось 5 слоёв оригинальных данных, зеленый – 10 слоёв соответственно. На графиках важно выделить начальные точки, где синтетические данные ещё не использовались. Отмечается существенный рост качества сегментации при добавлении 90 синтетических тайлов, что составляет около 15 % от общего числа данных в наборе, содержащем 5 реальных слоев, и 7,5 % – в наборе, содержащем 10 реальных слоев (рис. 3). Эти точки соответствуют количеству аксонов 177 и 193 (165 синтетических аксонов) соответственно (рис. 4). Наилучшее качество наблюдается при увеличении количества синтетических изображений до 210 (рис. 3). Это указывает на то, что увеличение количества данных улучшает способность модели различать аксоны на изображениях. После добавления более 210 синтетических изображений не наблюдается повышение качества.



Роисунок 3. Зависимость качества сегментации от количества синтетических изображений

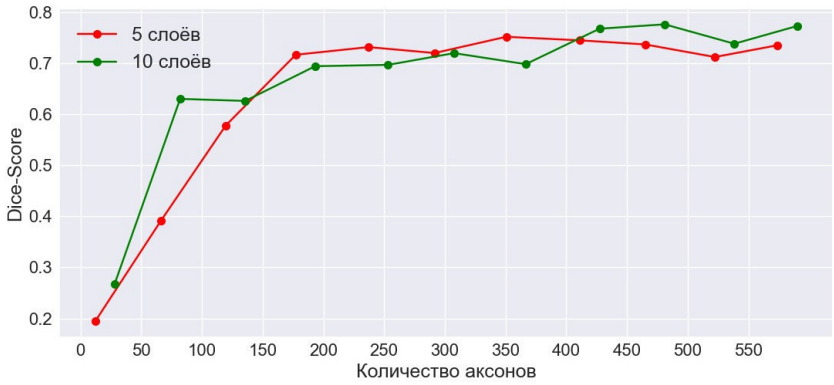


Рисунок 4. Зависимость качества сегментации от количества аксонов

Во время проведения эксперимента оказалось, что модель начинает путать постсинаптические уплотнения и аксоны, что заметно на рисунке 6. Это может происходить из-за отсутствия PSD в синтетических изображениях, поэтому во втором эксперименте они были добавлены при генерации.

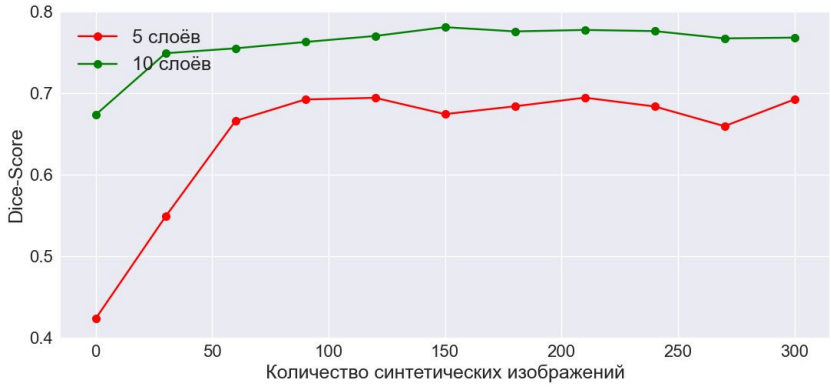


Рисунок 5. Зависимость качества сегментации PSD от количества добавленных данных с синтетическими аксонами

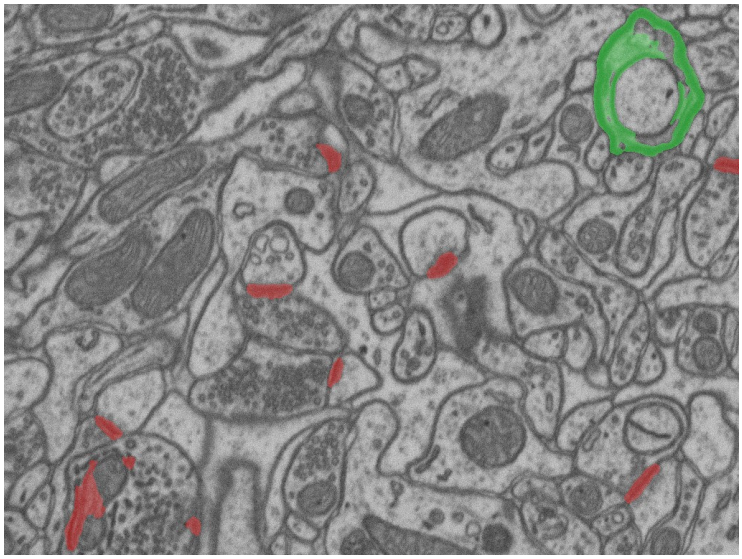


Рисунок 6. Маска сегментации аксонов. Зеленое наложение совпадает с разметкой, красное – ложная разметка

Второй эксперимент, изображенный на рисунках 7 и 8, также показывает положительную зависимость качества сегментации от количества добавленных синтетических изображений. Синий цвет отображает эксперимент, в котором использовались только синтетические данные. Красный цвет

отображает эксперимент, в котором использовались 5 слоёв оригинальных данных, остальные – синтетические, зеленый цвет – 10 слоёв оригинальных данных соответственно. На графиках важно выделить начальные точки, где синтетические данные ещё не использовались. Результаты довольно схожи с первым экспериментом: при добавлении 180 синтетических аксонов и больше качество сегментации Dice становится больше 0,74, но в отличие от первого эксперимента рост качества не прекращается на 0,78, а достигает 0,795 для красного графика при количестве аксонов 300 (рис. 8), что вдвое меньше, чем для первого эксперимента.

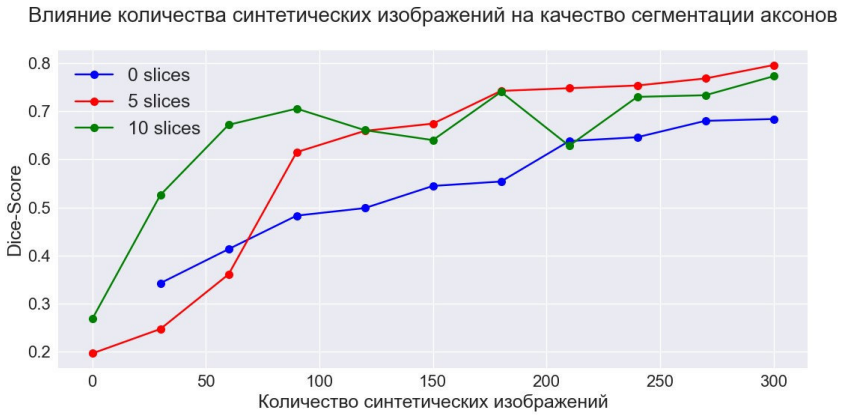


Рисунок 7. Зависимость качества сегментации от количества синтетических данных

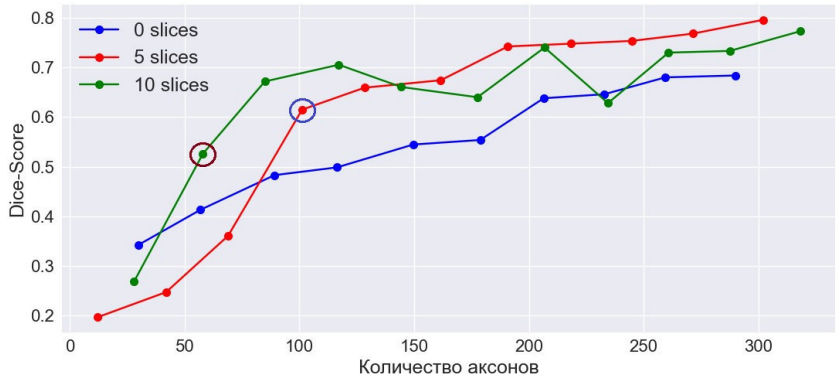
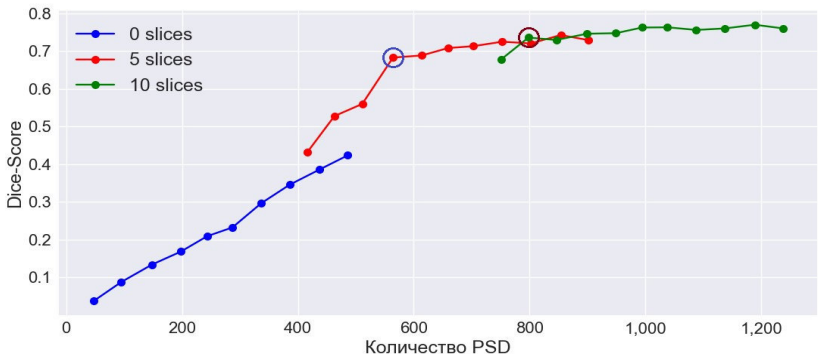


Рисунок 8. Зависимость качества сегментации от количества аксонов

На рисунке 9 видно, что качество сегментации PSD также растет и довольно быстро достигает уровня в 75 % в эксперименте с 10 оригинальными изображениями и 72 % с 5 оригинальными изображениями. В первом эксперименте с 5 оригинальными изображениями качество сегментации PSD меньше 70 %, однако с 10 изображениями качество сегментации достигает 77 %. Добавление PSD в синтетические данные не несет изменений в качестве их сегментации, однако аксоны определяются увереннее.



Рсиунок 9. Зависимость качества сегментации PSD от их количества

По результатам экспериментов заметно, что для обучения нейронной сети с синтетическими данными очень важно добавлять настоящие данные или генерировать синтетику, неотличимую от настоящих данных. Это позволяет улучшить качество сегментации. Синтетические данные, в свою очередь, могут быть полезны для увеличения объема обучающего набора и создания разнообразия, что особенно важно в случаях, когда реальные данные ограничены или труднодоступны.

Заключение

В рамках исследования на малой части датасета электронной микроскопии было проанализировано влияние добавленного объёма синтетических данных на качество сегментации, редко встречаемого в исходном наборе данных класса аксон. Полученные результаты показали, что использование синтетических данных может существенно повысить качество сегментации. При сравнении с исследованием [13], где использовались 42 слоя из того же датасета и 1000 синтетических тайлов, сгенерированных аналогичным способом и включающих все 5–6 классов, различия в качестве сегментации аксонов оказались незначительными. Это связано с сопоставимым количеством аксонов в итоговых выборках, на 1000 тайлов генерации приходится примерно 475 аксонов.

Несбалансированные наборы данных создают сложности в предсказаниях как для редких классов, так и для классов, схожих с ними, что приводит к неопределенности в предсказаниях. Мы продемонстрировали, что добавление простых синтетических изображений для редких классов существенно улучшает качество сегментации. Это особенно выражено при небольшом дополнении как для основного редкого класса аксон, так и для схожего по интенсивности класса PSD, что позволяет снизить путаницу между ними.

Источник финансирования

Работа выполнена при поддержке Министерства экономического развития Российской Федерации (соглашение о предоставлении гранта № 139-15-2025-004 от 17 апреля 2025 г., ИГК 000000Ц313925Р3Х0002).

Список литературы

1. Sergey I. Nikolenko. Synthetic Data for Deep Learning (2019). ArXiv preprint arXiv:1909.11512
2. Jonas Rabensteiner, Cynthia I. Ugwu, Oswald Lanz. Improving Semantic Segmentation Models through Synthetic Data Generation via Diffusion Models (2024). ICLR 2024. Workshop DMLR.
3. Daniel Saragih, Atsuhiro Hibi MSc, Pascal N. Tyrrell PhD (2024). Using Diffusion Models to Generate Synthetic Labeled Data for Medical Image Segmentation. arXiv preprint arXiv:2310.16794
4. V. P. Agostina J. Larrazabal, Nicol´as Nieto. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. PNAS, 2020.
5. EPFL Computer Vision Lab. Electron Microscopy Dataset.
6. ITMM, 6-Class Labels for EPFL EM Dataset (2023). GitHub repository. https://github.com/GraphLabEMproj.
7. N.A. Sokolov, E.P. Vasiliev, A.A. Getmanskaya. Generation and study of the synthetic brain electron microscopy dataset for segmentation purpose. Computer Optics, 2023.
8. GraphLab EM Project (2021). Synthetics: Synthetic Data Generator for EM Segmentation, GitHub repository https://github.com/GraphLabEMproj/Synthetics
9. Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation (2015). ArXiv preprint arXiv:1505.04597.
10. Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval. Medical Image Segmentation Review: The Success of U-Net. arXiv preprint arXiv:2211.14830 (2022).
11. A. A. Getmanskaya, Nikolai A. Sokolov, V. E. Turlapov. Multiclass U-Net Segmentation of Brain Electron Microscopy Data Using Original and Semi-Synthetic Training Datasets. Programming and Computer Software 48 (2022) 164–171.
12. Nikolai Sokolov. UnetClass: GitHub repository (2024). https://github.com/ NikolaySokolov152/UnetClass
13. Nikolay Sokolov, Alexandra Getmanskaya, Vadim Turlapov. AI Diffusion Model-Based Technology for Automating the Multi-Class Labeling of Electron Microscopy Datasets of Brain Cell Organelles for Their Augmentation and Synthetic Generation. Technologies, 2025.