Нейросетевые методы поиска и сегментации объектов в данных современных космических обзоров (eROSITA, ART-XC)

Научные руководители:

Герасимов С.В., к.ф.-м.н. Мещеряков А.В.

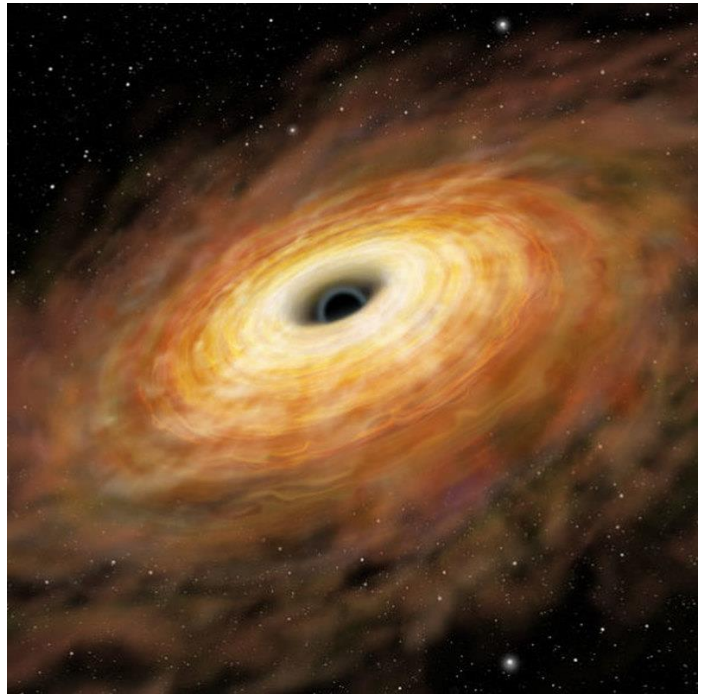
Студент:

Немешаева Алиса, 3 курс бакалавриата ВМК МГУ

Введение

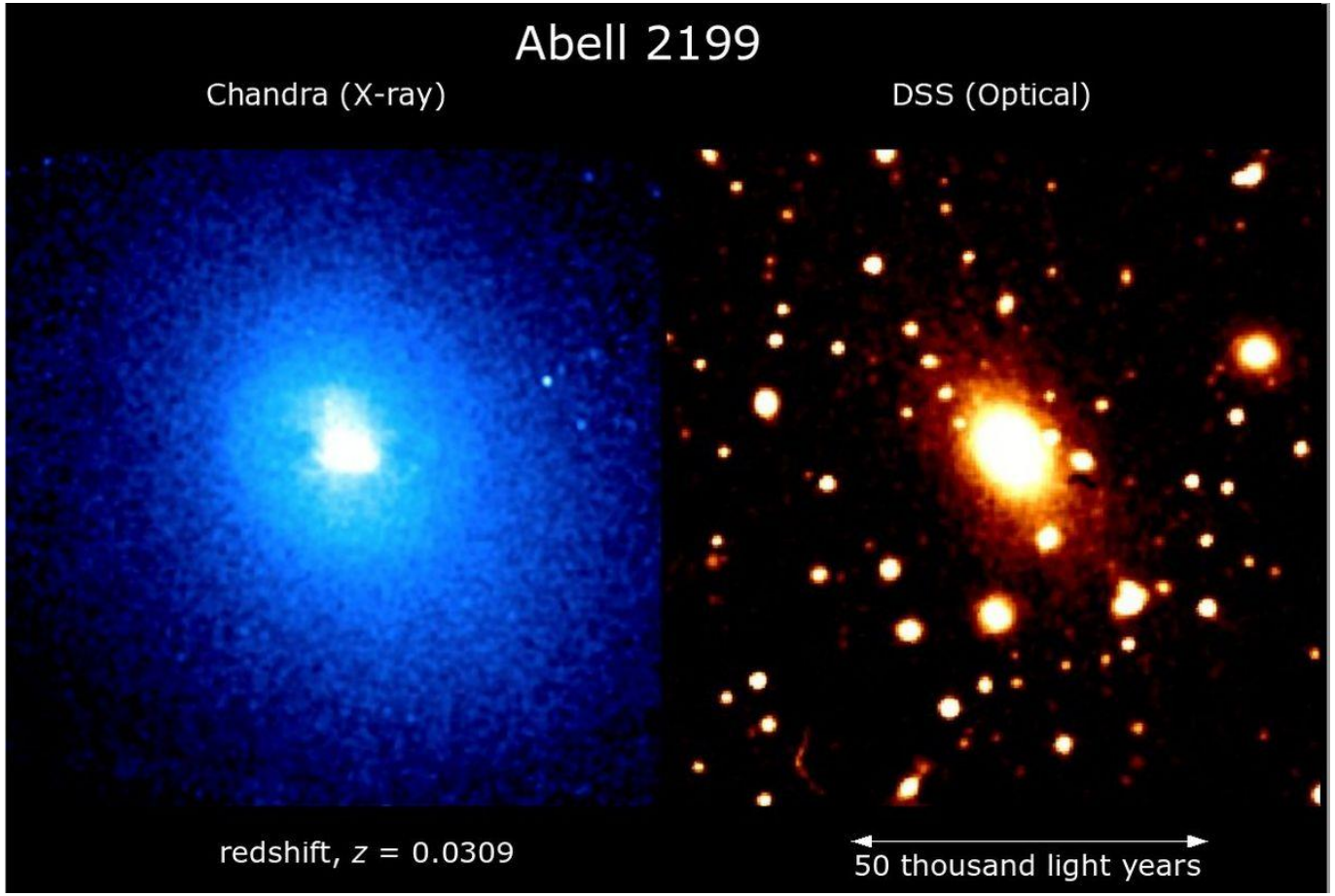
В 2019 году произошёл запуск космической обсерватории СРГ с телескопами eROSITA и ART-XC на борту. Основной задачей этих телескопов является создание обзора всего неба в рентгеновском диапазоне. Данные, полученные от этих телескопов будут использоваться для обнаружения трех типов объектов:

- Скопления галактик.
- Сверхмассивные чёрные дыры.
- Рентгеновские звёзды в галактике Млечный путь.



Введение

- Скопления - это гравитационно связанные системы, которые являются самыми большими структурами во Вселенной. Скопления галактик играют важную роль в задачах определения параметров Вселенной. Скопления галактик излучают энергию в разных диапазонах, и их можно наблюдать не только в рентгеновских данных.



Введение

- Полные обзоры неба, полученные телескопом eROSITA, появятся к июню 2020 года, поэтому на данный момент есть возможность подготовить модели для сегментации данных на примере других диапазонов.
- Будут использоваться данные оптического диапазона. В данной работе будут использоваться данные телескопа Pan-STARRS1.
- Данные этого телескопа покрывают ту часть неба, что станет доступна в рентгеновском диапазоне после завершения исследований с помощью eROSITA.
- Кроме того, обзор этого телескопа отличается максимальной глубиной среди других оптических телескопов и его данные находятся в общем доступе.

Телескоп Pan-STARRS 1

Актуальность

В последние годы методы глубокого обучения стали играть важную роль в анализе данных. Нейросетевые модели показывают высокие результаты в области компьютерного зрения и в частности в задачах сегментации и детекции. Всё более часто они применяются и для решения задач астрофизики. Методы глубокого обучения дают много преимуществ при анализе данных:

- Стандартные алгоритмы сегментации усредняют информацию по нескольким каналам, в то время как с помощью нейросети можно охватить данные полностью. Таким образом, нейросеть будет получать «сырые» данные, что экономит время и исключает необходимость контролировать процесс предобработки данных. Кроме того, нейросеть, используя все данные, может получить информацию о калибровке телескопа прямо из обзоров.
- Аналогичнные методы можно использовать для сегментации одновременно разнородных данных.
- Каждый из классических методов имеет свои достоинства и недостатки, и для каждого диапазона излучения существуют свои алгоритмы, в то время как нейросеть может стать универсальным средством для сегментации.

Постановка задачи

- Исследование проблемы сегментации данных в многоволновых диапазонах и разработка алгоритма для предобработки данных, а также создание нейросетевой модели, позволяющей выполнять сегментацию и детекцию скоплений галактик на оптических данных телескопа PS1.

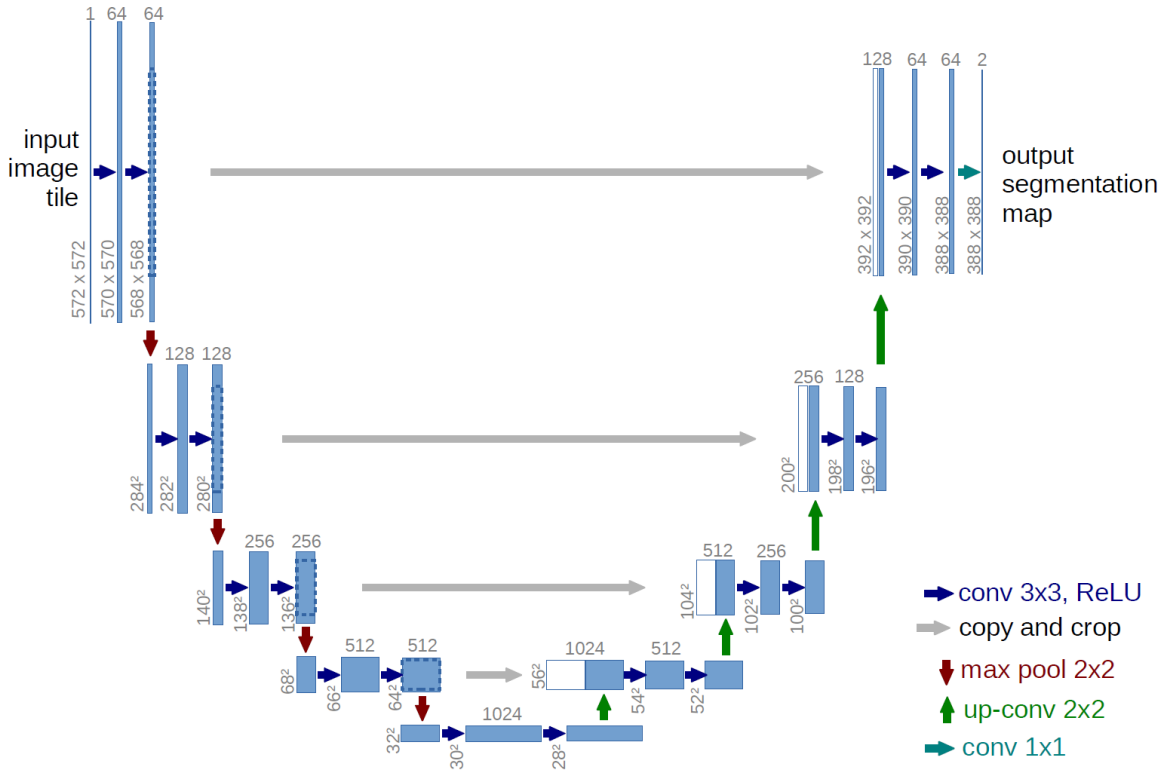
План задачи

Для достижения цели задание было разделено на несколько шагов:

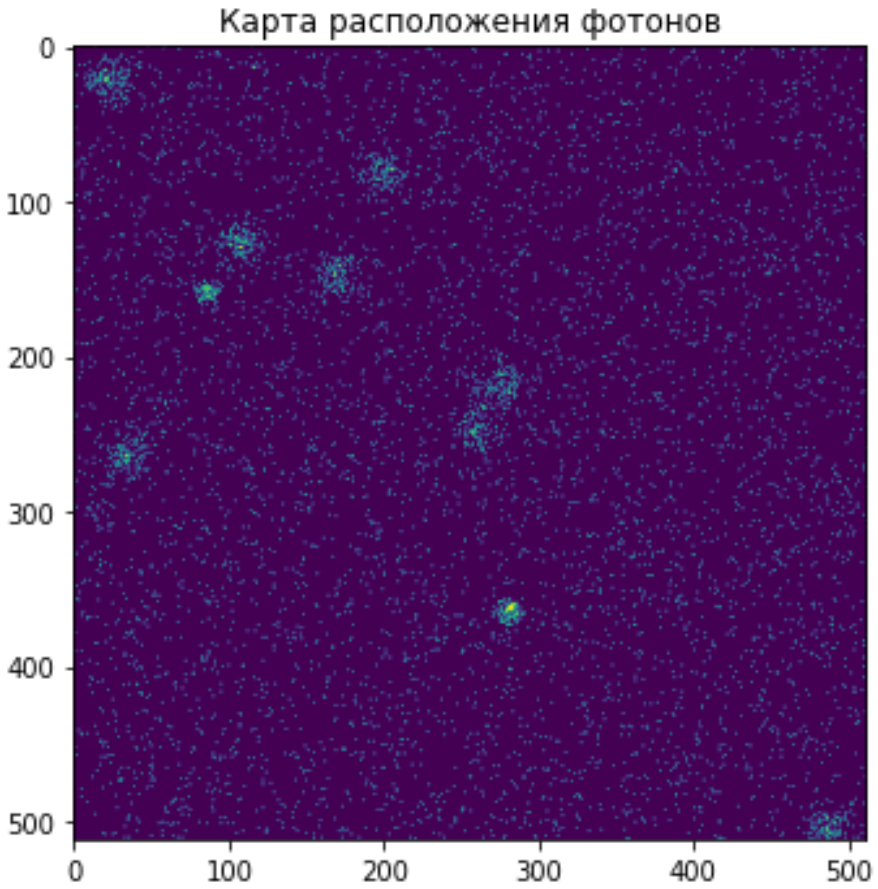
- Обзор существующих решений.
- Создание простейших симуляций рентгеновских данных и образца модели U-net и проверка работы U-net на данных симуляций.
- Создание модели
  - Предобработка.
    - Загрузка и обработка данных о скоплениях.
    - Генерация «патчей» и загрузка обзоров неба PanSTARRS1 из области патчей.
    - Преобразование данных PanSTARRS1 в двумерные матрицы для загрузки в нейросеть.
  - *Обучение модели, подбор параметров модели (количество слоёв, методы аугментации, размер батча, количество эпох обучения)*
  - *Тестирование модели на заранее выбранных данных.*

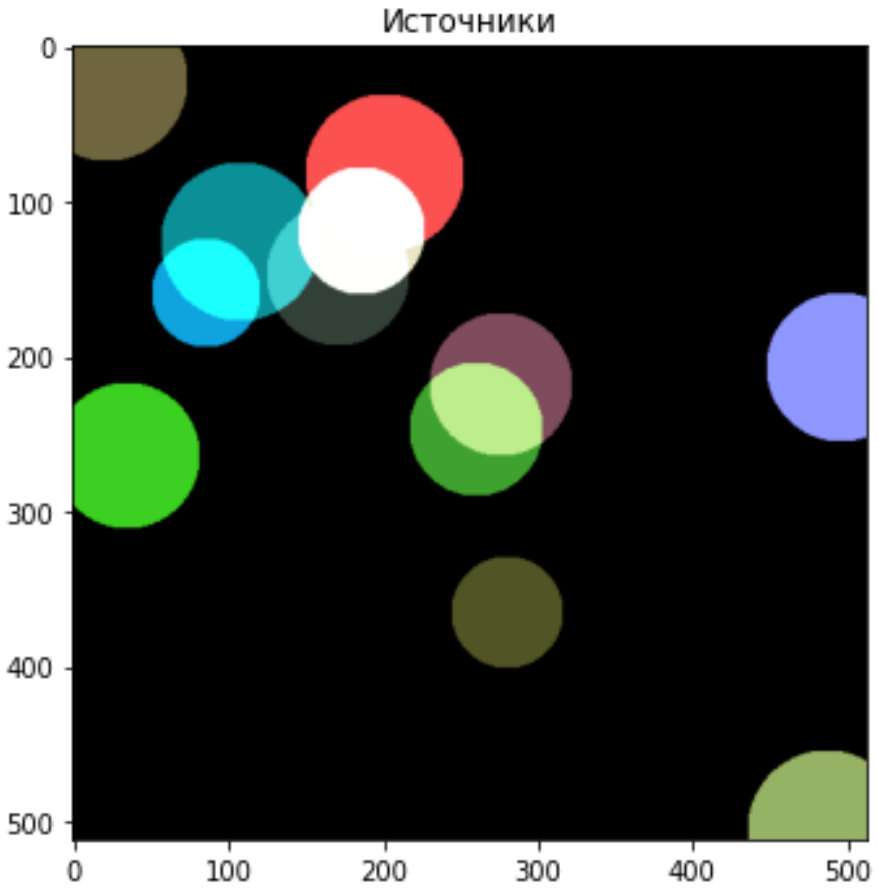
Обзор. U-net

- U-net является стандартной архитектурой для сегментации данных. Она подходит для проверки идеи использования методов глубокого обучения для сегментации скоплений. Её симметричная структура позволяет абстрагировать данные изображения, подаваемого на вход, в то время как skip-connection слои помогают увеличивать точность сегментации.



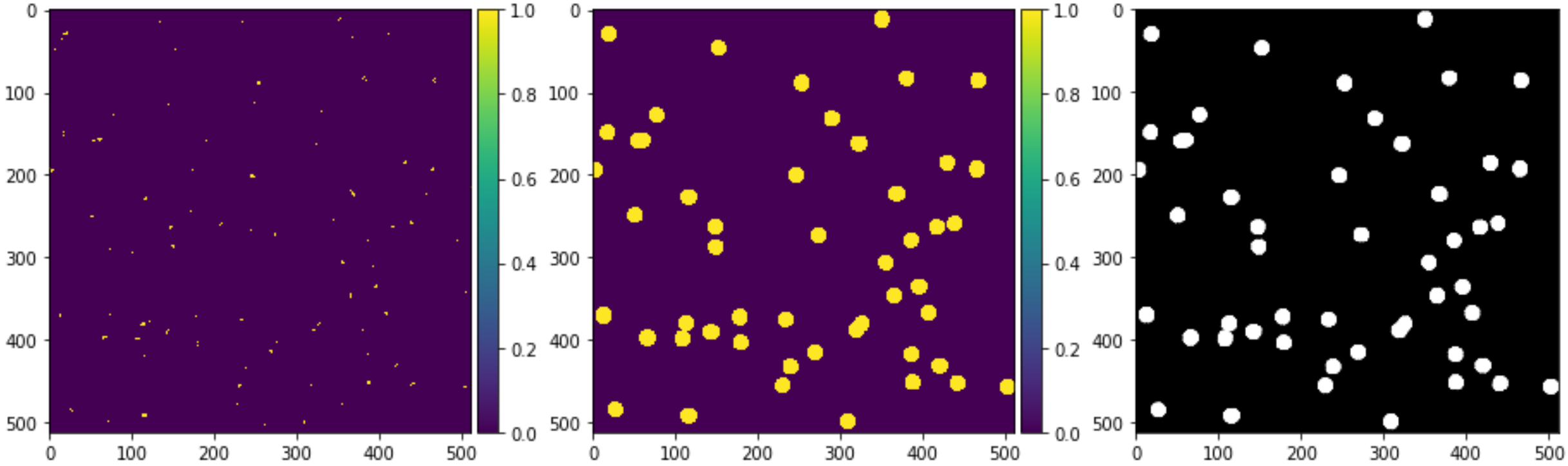
Описание практической части. Симуляция и первый образец нейросети





Пример генерации данных и масок для каждого скопления

Описание практической части. Симуляция и первый образец нейросети

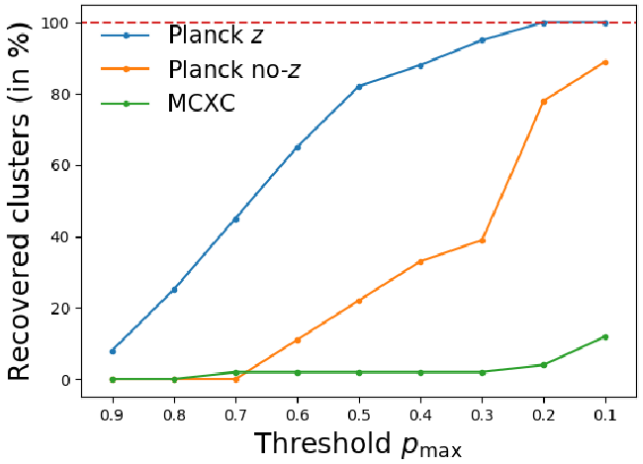


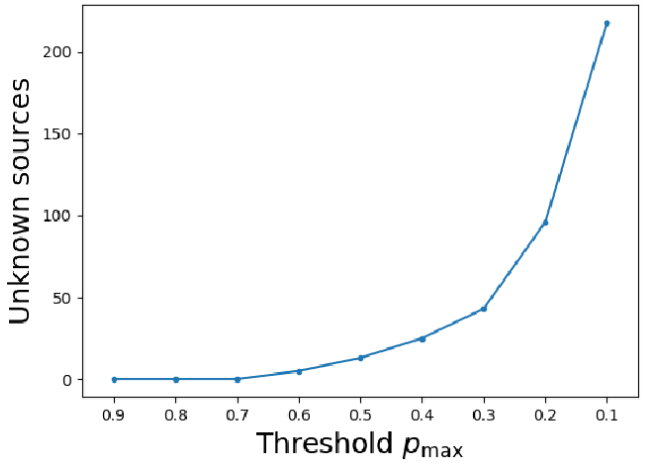
Пример работы симуляции и пример сегментированного изображения. Итоговая точность сегментации составляет 0.9978 для симулированных данных.

Обзор существующих решений

В первую очередь рассмотрим работу о детекции эффекта Сюняева-Зельдовича. Её автор тоже использует для сегментации данных архитектуру U-net. Основной целью описываемой работы являлось создание алгоритма для детекции источников через эффект Сюняева-Зельдовича по данным телескопа «Планк».
Кроме самих обзоров неба, полученных «Планком», использовались еще три каталога скоплений для создания целевых данных:

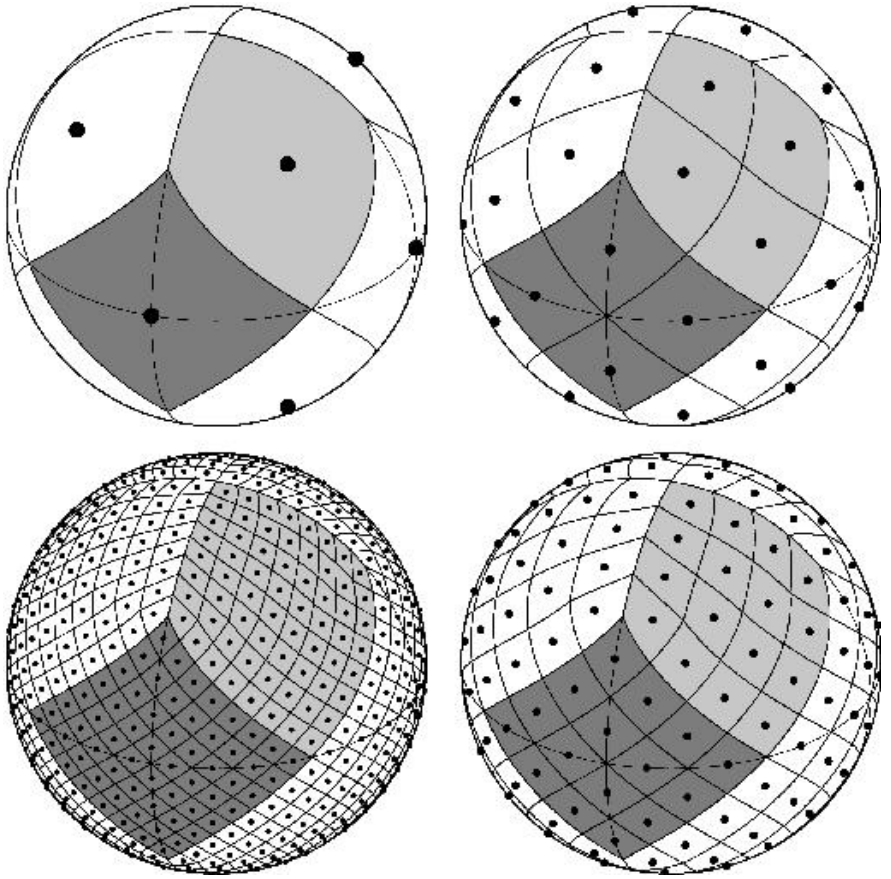
- PSZ2
- MCXC
- RedMaPPer





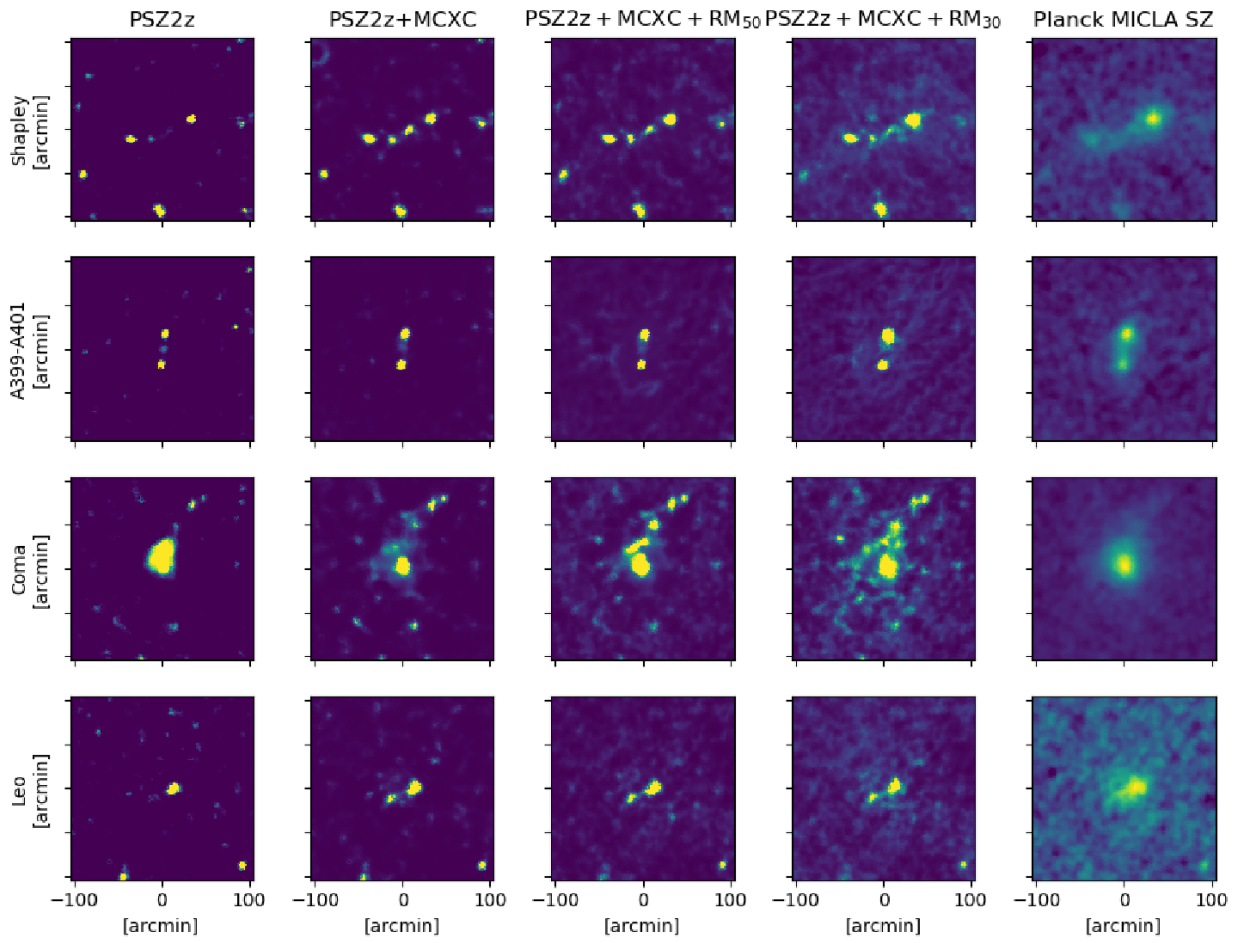
Обзор существующих решений

- В описываемой работе для создания тренировочных выборок использовалось разбиение неба проекцией HEALPix. Разбиение с параметром nside=2 позволяет получить 48 больших областей неба. Некоторые из них были использованы для тестирования полученной модели и для валидации, все остальные были использованы для обучения модели.



Обзор существующих решений

- Случайным образом в соответствующих областях разбиения выбирались центры патчей и их ориентации для создания тренировочных и тестовых выборок. Каждый патч представлял из себя изображение размера 64 x 64 с шестью каналами различных данных.
- После этого 100000 патчей были использованы для обучения нейросети. Обучение длилось более 30 эпох
- Как можно увидеть из результатов, лучше всего нейросеть сегментирует в данных телескопа «Планк» скопления из каталога того же самого «Планка».
- В описываемой статье использовалась классическая версия архитектуры U-net.



Практическая часть. Предобработка данных.

В качестве списка скоплений будем использовать каталог PSZ2. Аналогичным образом будем генерировать патчи для создания тренировочных и тестовых выборок. Патчи выбирались так, чтобы в их окрестности находилось хотя бы одно из скоплений нужного каталога. После этого в базе данных PS1 запрашивался список объектов, подходящих под заданные условия.

Для каждого объекта из PS1 загружались следующие данные:

- id объекта.
- Координаты объекта.
- (filter)PSFFlux - информация о ядре объекта.
- (filter)PSFFluxErr - ошибка измерения (filter)PSFFlux.
- (filter)KronFlux - информация о полном свете объекта.

Из-за особенностей данных, некоторые объекты могут повторяться (то есть их координаты полностью совпадают). В таких ситуациях нужно провести слияение объектов --- для каждого из параметров (filter)PSFFlux и (filter)KronFlux нужно сохранить значение из той строки, где ошибка соответствующего измерения будет ниже.

После того, как будут получены данные для обучения, их нужно из таблиц преобразовать в двухмерные матрицы изображений, чтобы создать выборки с количеством каналов, совпадающих с количеством исследуемых параметров у объектов.

Текущие результаты

На данный момент было решено несколько подзадач, поставленных для решения общей проблемы:

- Созданы простейшие программы для генерирования симуляций в рентгеновском диапазоне.
- Проверена работоспособность модели U-net на данных симуляций.
- Обзор методов сегментирования.
- Обработаны каталоги источников.
- Создан код для генерирования «патчей» для обучения и тестирования нейросети.
- Начата работа по обработке данных PS1.