МГУ им. М. В. Ломоносова, факультет ВМК

Курсовая работа

на тему: «Нейросетевые методы поиска и сегментации объектов в данных современных космических обзоров (eROSITA, ART-XC)»

Студент гр. 320
Немешаева Алиса
Научные руководители:
Герасимов С. В., к.ф-м.н Мещеряков А. В

Москва 2020

Содержание

1 Аннотация

Данная работа рассматривает возможность применения нейросетевых методов к решению проблемы сегментации и детекции объектов по многоволновым данным космических телескопов (в данном случае оптического, инфракрасного и рентгеновского диапазонов). В качестве основы для нейросетевой архитектуры использовалась модель U-net.

2 Введение

2.1 Скопления галактик

В 2019 году произошел запуск космической обсерватории СРГ (Спектр-Рентген-Гамма) 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 астрономических объектов трёх категорий:

а) Скопления галактик.
б) Сверхмассивные чёрные дыры.
в) Рентгеновские звёзды в галактике Млечный путь.

Наибольший интерес представляют скопления галактик. Скопления — это гравитационно связанные системы, которые являются самыми большими динамически связанными структурами во Вселенной. Скопления галактик играют важную роль в задачах определения космологических параметров Вселенной. Например, зная расстояние до скопления и его красное смещение (параметр, по которому можно понять, как объект отдаляется от наблюдателя), можно уточнить постоянную Хаббла, входящую в закон Хаббла, который описывает скорость расширения Вселенной. Кроме того, соотношение компонент материи в скоплениях должно отражать средний состав Вселенной, что позволяет измерить вклад барионов в общую плотность Вселенной.

Скопления галактик излучают энергию в разных диапазонах, и их можно наблюдать не только в рентгеновских данных. Однако рентгеновские данные являются лучшим источником информации о скоплениях, так как в оптическом диапазоне наблюдается не только излучение далёких скоплений, но и свет ближайших звёзд из Млечного пути. Несмотря на это, исследование оптического диапазона тоже может принести результаты, так как можно исследовать области неба, где излучение Млечного пути сравнительно меньше. К тому же, можно использовать несколько спектров одновременно, увеличивая точность сегментации.

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.

В первую очередь будут использоваться данные оптического диапазона. Видимое излучение — тот диапазон частот, что доступен глазу человека. На текущий момент существует большое количество оптических телескопов, и, как следствие,

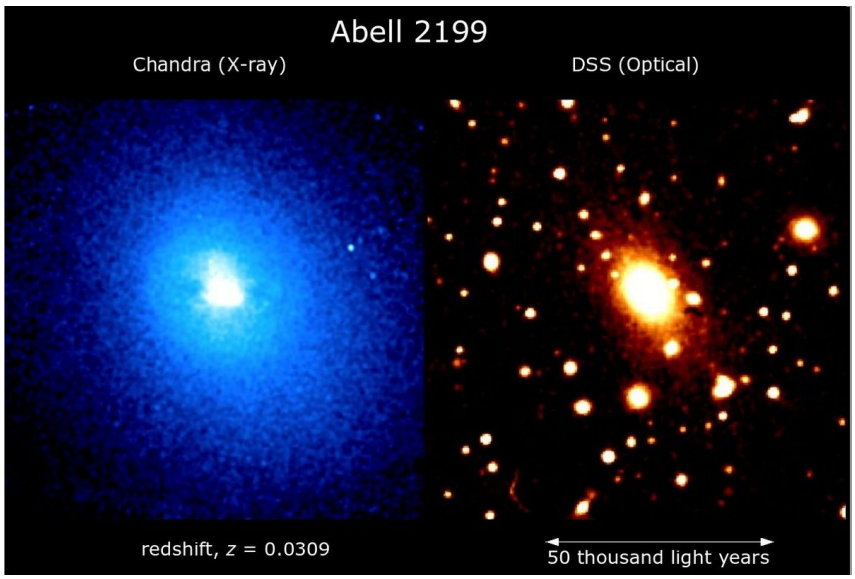


Рисунок 2.1 — Скопление «Абель 2199» в оптическом и рентгеновском диапазонах. [1]

большое количество данных, извлеченных с их помощью. В данной работе будут использоваться данные телескопа Pan-STARRS1, который является частью системы телескопов Pan-STARRS (Panoramic Survey Telescope and Rapid Response System). Этот телескоп построен на вершине гавайского вулкана Халеакала. На 2007 год он обладал самой большой светочувствительной матрицей в мире. Кроме того, его данные находятся в общем доступе [2].

2.2 Нейросетевые методы

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. Характеристики телескопа eROSITA позволят получить рентгеновские данные очень высокого качества (то есть с низким количеством шума), и методы глубокого обучения дают много преимуществ при анализе данных:

а) Стандартные алгоритмы сегментации усредняют информацию по нескольким каналам, в то время как с помощью нейросети можно охватить данные полностью и исследовать вопрос с новой стороны. 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, что невозможно сделать при использовании классических методов.

б) Аналогичнные методы можно использовать для сегментации одновременно разнородных данных. То есть для улучшения качества сегментации можно исследовать параллельно разные диапазоны частот и находить взаимосвязь между разными спектрами.

в) 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.

U-net [3] является стандартной архитектурой для сегментации данных. Она идеально 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.

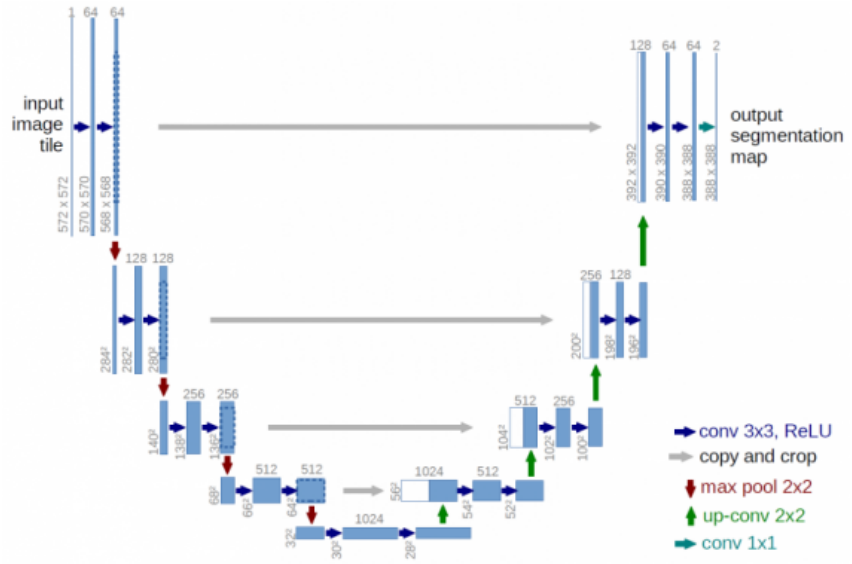


Рисунок 2.2 — Структура модели U-net [3]

3 Постановка задачи

На данный момент имеется доступ к оптическим данным, полученным с помошью телескопа PS1. Выбрана архитектура нейросетевой модели для сегментации изображений. Требуется обработать данные PS1, преобразовать их в подходящий для нейросети вид, и получить алгоритм сегментации скоплений галактик.

В более формальном и подробном виде задачу можно описать так: при имеющихся обработанных данных, полученных из космических обзоров, требуется получить матрицы сегментации, где для каждого пикселя матрицы мы будем иметь информацию о вероятности, с которой в данном пикселе находится скопление. Впоследствии координаты пикселей изображений можно преобразовать в небесные координаты способом, обратным к способу, с помощью которого были получены изображения космических обзоров.

Для достижения цели задание было разделено на несколько шагов:

а) Создание простейших симуляций рентгеновских данных.

б) Создание образца модели U-net.

в) Проверка работы U-net на данных симуляций.

г) Загрузка и обработка данных о скоплениях.

д) Генерация «патчей» — небольших областей неба, на которых будет тренироваться нейросеть.

е) Загрузка и обработка обзоров неба PanSTARRS1 из области патчей.

ж) Преобразование данных PanSTARRS1 в двумерные матрицы для загрузки в нейросеть.

з) Обучение модели, подбор параметров модели (количество слоёв, методы аугментации, размер батча, количество эпох обучения)

и) Тестирование модели на заранее выбранных данных.

к) Преобразование масок сегментации в координаты (детектирование скоплений).

л) Сравнение полученных скоплений с существующими каталогами.

Пока что не существует какого-то универсального метода для сегментации и детекции скоплений на оптических данных, и есть возможность применить методы глубокого обучения в данной области.

Кроме того, сами по себе данные телескопов не похожи на обычные изображения, которые хранятся в матрицах из одного или трех фильтров и используют однобайтовые целые значения, не требующие большого количества памяти. Матри-

цы, получающиеся при преобразовании данных из космических координат, обычно получаются разреженными и для них приходится использовать значения с плавающей точкой. Более того, нужно учитывать точность преобразования и выбирать достаточно детализированное разбиение проекции на пиксели изображения, иначе разные объекты могут слиться в один. Угловой размер области тоже имеет значение, так как на слишком маленьких частях неба искать такие большие объекты, как скопления галактик, бессмысленно.

4 Обзор существующих решений рассматриваемой задачи или её модификаций

4.1 Сегментация и детекция данных микроволнового диапазона

В первую очередь рассмотрим работу о детекции эффекта Сюняева-Зельдовича [4] в микроволновых данных. 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». Соответственно, кроме самих обзоров неба, полученных «Планком», использовались еще три каталога скоплений для создания целевых данных:

а) PSZ2. Этот каталог был получен по данным «Планка» при помощи алгоритмов согласованного мультифильтра и PowellSnakes.

б) MCXC (Meta-Catalogue of X-ray detected Clusters). Это объединенный каталог из всех других каталогов скоплений, полученных из данных телескопа ROSAT.

в) RedMaPPer (Red-sequence Matched-filter Probabilistic Percolation). Каталог скоплений, полученный с помощью одноимённого алгоритма из данных оптического диапазона.

Несмотря на то, что в этих каталогах содержатся данные об объектах, детектированных в разных диапазонах, они содержат довольно большое количество общих объектов (поэтому при создании тренировочных выборок нужно удалять из каталогов повторы).

В описываемой работе для создания тренировочных выборок использовалось разбиение неба проекцией HEALPix (Hierarchical Equal Area isoLatitude Pixelisation).

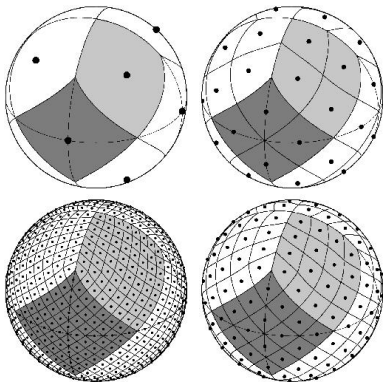


Рисунок 4.1 — Примеры разбиения сферы HEALPix [5]

Разбиение с параметром $n_{side} = 2$ 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.

Случайным образом в соответствуюхих областях разбиения HEALPix выбирались центры патчей и их ориентации для создания тренировочных, валидационных и тестовых выборок. Каждый патч представлял из себя изображение размера 64 x 64 с шестью каналами различных данных. Размер каждого пикселя на таких патчах составлял 1.7 arcmin.

После этого 100000 патчей были использованы для обучения нейросети. Обучение длилось более 30 эпох.

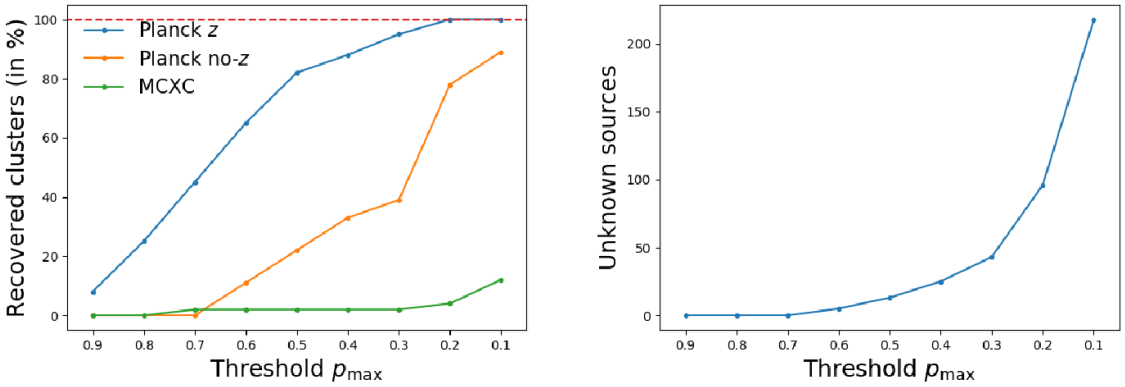


Рисунок 4.2 — Результаты исследования работы [4]

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: каждый из блоков кодировщика состоит из двух свёрток с ядром 3 x 3 с последующими активациями ReLU после каждой свёртки за ними следует слой MaxPooling для уменьшения размерности входных данных. Всего в кодировщике пять таких блоков. Блоки декодировщика состоят из слоя Upsampling, повышающего размерность предыдущего слоя, конкатенации выхода из кодировщика с соответсвующими размерами и так же двух сверток с ReLU. Кроме того, после каждого слоя свёртки добавлен слой Dropout с параметром 0.2 для предотвращения переобучения. После всех блоков декодировщика добавляется слой активации сигмоиды для последующего использования кросс-энтропии как loss-функции. Количество фильтров для первого блока варьировалось от 8 до 128.

Для детекции скоплений на полученных из нейросети данных, зонами скоплений обозначались зоны, занимающие пиксели со значением больше p_{max}, а затем на них находились барицентры, которые впоследствии считались предсказанными центрами скоплений. Таким образом большая часть скоплений из каталога PSZ2 была успешно распознана нейросетью.

4.2 Сегментация и детекция данных рентгеновского диапазона

На данный момент существует несколько пакетов программ, с помощью которых осуществляется сегментация данных, полученных с помощью рентгеновских телескопов. Одним из них является CIAO (CHANDRA INTERACTIVE ANALYSIS OF OBSERVATIONS), разработанный специально для Космической рентгеновской обсерватории «Чандра».

— *celldetect* использует свёртку с изменяющимся размером ядра для сегментации. Для каждого положения клетки вычисляется отношение сигнал/шум, сравнивается со средним значением для фона, и если оно больше, то регистрируется источник, в котором может находится объект.

— *vtpdetect* использует обнаружение источников с помощью тесселяции и перколяции Вороного (VTP) для определения индивидуальных плотностей или потоков для каждого данного пикселя. Затем инструмент анализирует распределение плотностей для обнаружения объектов.

— *wavdetect* использует свёртку вэйвлет функции с астрономических изображением. Такое преобразование позволяет удалить с изображения фон и локализовать структуры, размеры которых близки к масштабу вейвлет-преобразования.

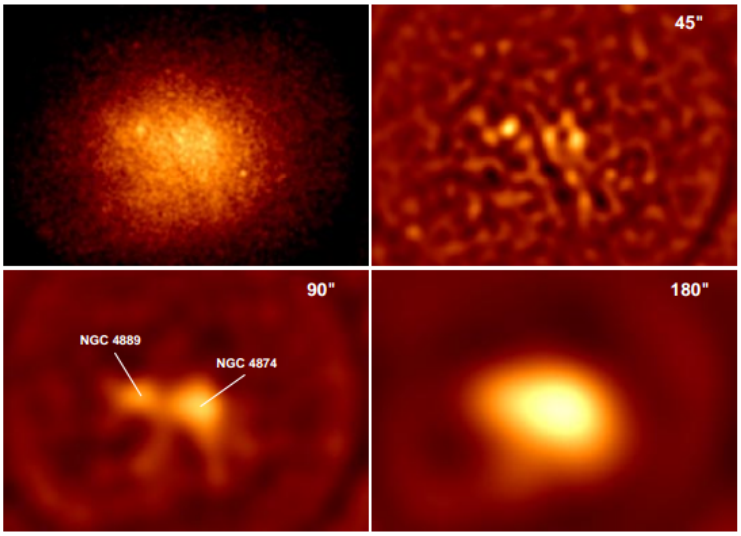


Рисунок 4.3 — Вэйвлет-анализ скопления Кома. Сырое изображение и изображение с применёнными к нему вэйвлет-преобразованиями с разным масштабом [6]

5 Исследование и построение решения задачи

5.1 Создание симуляций

Для создания симуляций использовался простейший алгоритм с использованием генерации случайных данных с помощью равномерного распределения, распределения Пуассона и распределения Гаусса. Поэтапно создание симулированных данных выглядит так:

а) Равномерным распределением создаём заданное количество скоплений на изображении.

б) Для каждого скопления выбираем размер и количество объектов из распределения Пуассона.

в) Для каждого скопления координаты объектов генерируются из распределения Гаусса.

г) Равномерно генерируются координаты объектов шума.

5.2 Формирование данных для обучения

Так же, как и в статье [4], в качестве списка скоплений будем использовать каталог PSZ2. Аналогичным образом будем генерировать патчи для создания тренировочных, валидационных и тестовых выборок (в том числе для валидации и тестирования будут выбраны те же пиксели разбиения $n_{side} = 2$. Патчи выбирались так, чтобы в их окрестности находился хотя бы одно из скоплений нужного каталога. После этого в базе данных PS1 (Pan-STARRS1) запрашивался список объектов, подходящих под заданные условия.

Для каждого объекта из PS1 загружелись следующие данные:

а) id объекта.
б) Координаты объекта.
в) (filter)PSFFlux - информация о ядре объекта.
г) (filter)PSFFluxErr - ошибка измерения (filter)PSFFlux.
д) (filter)KronFlux - информация о полном свете объекта.
е) (filter)KronFluxErr - ошибка измерения (filter)KronFlux.

Вместо (filter) подставляется одна из букв {g, r, i, z, y}, обозначающих фильтр, которым обрабатывалась информация об объектах. Таким образом, информация об излучении объекта хранится в 20 параметрах, включая ошибки.

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. Оставшиеся значения ошибок можно либо удалить из данных, либо использовать их для аугментации: каждый раз при переносе значения какого-либо параметра, прибавлять к нему случайное значение из нормального распределения $N(0, (filter)(parameter)Err)$.

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.

Для примера, таблица с данными для 50 патчей содержит около 10 миллионов объектов. Обработка сотни объектов будет длится несколько минут, поэтому необходимо отдельно распараллелить процесс.

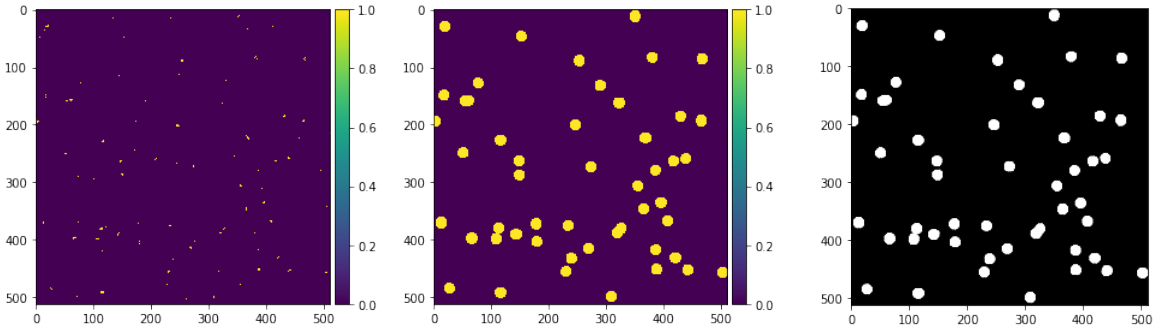
Для обучения планируется использование параметров (filter)KronFlux, то есть у первой версии нейросетевой модели будет 5 входных слоёв. Значения (filter)PSFFlux нужны для того, чтобы заменить ими отсутствующие измерения (filter)KronFlux для некоторых объектов.

6 Описание практической части

6.1 Симуляции и первый образец модели

При подготовке к созданию итоговой модели в первую очередь создавались симуляции (искуственные данные, похожие по статистическим распределениям на настоящие, но по своей структуре более простые).

После этого на созданных симуляциями данных тренировались первые образцы нейросетевых моделей. Общая архитектура схожа с моделью из статьи [4], но в данном случае вместо слоёв Dropout использовались слои батч-нормализации и вместо 5 блоков было добавлено 3 блока с 32 фильтрами в первом блоке.



a) b) c)

Рисунок 6.1 — a) Данные объектов из симуляций; b) Данные источников из симуляций; c) Результат работы нейросети;

Выше приведен пример работы симуляции а также пример сегментированного изображения. Итоговая точность сегментации составляет 0.9978 для симулированных данных.

6.2 Обработка каталога скоплений

Далее, начинается обработка настоящих данных. Нужно загрузить и обработать каталог скоплений Planck, который будет разделен на два подкаталога:

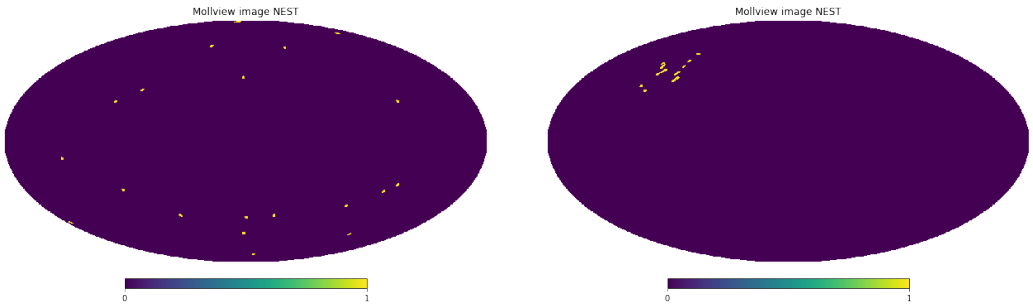
а) planck_z (скопления с измеренным красным сдвигом)
б) planck_no_z (скопления без информации о красном сдвиге)

6.3 Обработка данных PS1

Поскольку поиск в базе данных PS1 можно вести только в системе координат ICRS, центры скоплений нужно преобразовать в нужный вид. Однако проекция

healpix будет использовать галактические координаты, поэтому координаты найденных объектов будут сохранены соответственно в галактической системе.

После того, как были получены данные по скоплениям, можно начать загрузку и обработку данных из обзоров PS1. В каждом пикселе из разбиения с $n_{side} = 2$ генерируется определенное количество патчей. Центры патчей выбираются случайным образом как пиксели, рядом с которыми есть скопления. Размер каждого патча задан как 2048 x 2048 в разбиении с $n_{side} = 2^{17}$, и так как пиксели HEALPix могут иметь протяжённую структуру, итоговый радиус патча вычислялся как расстояние от центра патча до дальнего угла для патча размером 2050 x 2050. В итоге максимальный радиус оказался равен $\approx 0.9°$. Для преобразований использовалась библиотека healpy.



a) b)

Рисунок 6.2 — Сгенерированные центры патчей: a) Для всего неба; b) Для пикселя №6 из разбиения $n_{side} = 2$;

Далее данные нужно преобразовать в формат двумерных матриц для загрузки в нейросеть, перед этим вычислив для каждого объекта номер пикселя healpix, к которому он относится. Объекты, относящиеся к одним и тем же пикселям, нужно соответствующим образом отождествить, сохранив те, для которых значения ошибок меньше.

Для этого создавались матрицы соответсвий между координатами изображения и номерами пикселей из healpix-разбиения. Перед началом создания такой матрицы мы всегда знаем номер пикселя для центра изображения, в нашем случае для координат $(1023, 1023)$. Библиотека healpy позволяет находить 8 соседних пикселей, зная вектор координат для заданного пикселя, поэтому постепенно, начиная с центра, можно найти все пиксели изображения.

После этого по построенным матрицам нужно каждый объект перенести на изображение. Это можно сделать двумя способами:

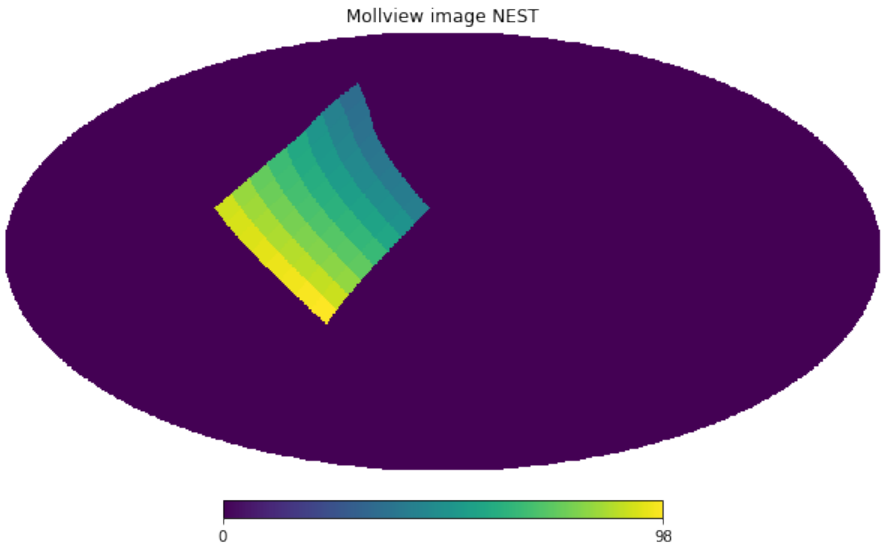


Рисунок 6.3 — Пример расположения двумерной матрицы на проекции неба для большего разбиения HEALPix

— Для яркости объекта можно на изображение на полученные координаты нанести круг с радиусом, пропорциональным яркости. Тогда матрица будет состоять из нулей и единиц. Такой метод подходит для визуальной оценки правильности преобразования координат.

— В заданные координаты записать значение нужного параметра. Такие матрицы и будут отправлены в нейросеть на обучение. Состоят из значений с плавающей точкой.

Точно так же перенести на двумерное изображение нужно и маски, определяющие наличие скоплений в данной области. Для каждого скопления вычисляются номера пикселей в радиусе $5'$, после чего они с помощью той же матрицы соответствий переносятся на изображение.

Заключение

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

а) Созданы простейшие программы для генерирования симуляций в рентгеновском диапазоне.

б) Проверена работоспособность модели U-net на данных симуляций.

в) Обработаны каталоги источников.

г) Создан код для генерирования «патчей» для обучения и тестирования нейросети.

д) Начата работа по обработке данных PS1.

Обработка данных PS1 требует больше времени, чем предполагалось, поэтому работа по данной проблеме будет продолжаться далее.

Подробная реализация перечисленных подзадач находится в публичном репозитории этой работы.

Список использованных источников

1. Abell 2199. — Wikipedia.

2. PanSTARRS1 (PS1) Catalog Archive Server Jobs System (CasJobs) service.

3. *O. Ronneberger P. Fischer, T. Brox.* U-Net: Convolutional Networks for Biomedical Image Segmentation / T. Brox O. Ronneberger, P. Fischer. — 2015.

4. *Bonjean, V.* Deep learning for Sunyaev–Zel'dovich detection in Planck / V. Bonjean. — Astronomy & Astrophysics, 2020.

5. HEALPix.

6. *Вихлинин, А. А.* Наблюдательная космология и изучение межгалактической среды по рентгеновским данным о скоплениях галактик / А. А. Вихлинин. — Российская Академия Наук, Институт Космических Исследований, 2002.