Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

Немешаева Алиса Алексеевна

Нейросетевые методы поиска и сегментации объектов в данных современных космических обзоров

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научные руководители:
Герасимов С.В.,

к.ф.-м.н. Мещеряков А.В.

Москва, 2020

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, микроволнового и рентгеновского диапазонов). В качестве основы для нейросетевой архитектуры использовалась модель U-net. В итоге были реализованы алгоритмы по предварительной обработке данных, обучена модель на основе архитектуры U-net на данных микроволнового обзора Planck и созданы два каталога детектированных объектов.

Содержание

1 Аннотация 0
2 Введение 3
2.1 Оптический диапазон 3
2.2 Рентгеновский диапазон 4
2.3 Микроволновой диапазон 5
2.4 Актуальность 6
3 Постановка задачи 8
4 Обзор существующих решений 9
4.1 Методы детекции скоплений в данных Planck 9
4.1.1 Метод согласованного мультифильтра: MMF1 и MMF3 9
4.1.2 PowellSnakes (PwS) 10
4.1.3 Нейросетевые подходы 11
4.2 Детекция скоплений по эффекту Сюняева-Зельдовича в данных SPT 11
4.3 Детекция скоплений по эффекту Сюняева-Зельдовича в данных ACT 11
5 Исследование и решение задачи 12
5.1 Обзор данных 12
5.1.1 Данные Planck 12
5.1.2 Каталоги скоплений по данным микроволнового диапазона 13
5.1.2.1 Каталог PSZ2 13
5.1.2.2 Каталог ACT 13
5.1.3 Каталоги скоплений по данным оптического диапазона: redMaPPer 14
5.1.4 Рентгеновкие каталоги скоплений 14
5.1.4.1 MCXC 14
5.1.4.2 2RXS 15
5.2 U-net 15
5.3 Другие нейросетевые модели 15
5.3.1 PSPNet 15

5.3.2 LinkNet 17
5.3.3 W-Net 17
6 Практическая часть 21
6.1 Модель 21
6.2 Предобработка данных и обучение модели 21
6.2.1 Нормализация данных 21
6.2.2 Каталоги для обучения 21
6.2.3 Разбиение данных на выборки 22
6.2.4 Генерация патчей и масок 23
6.2.5 Обучение модели 23
6.2.6 Метрики 23
6.3 Детекция скоплений на карте Planck. 25
6.4 Формирование каталога скоплений 27
6.5 Выбор эпохи для сканирования 28
Заключение 31
Список использованных источников 34

2 Введение

Уже в начале XX века было известно, что галактики формируются в скопления и сверхскопления галактик. Галактики создают собой нити и стены, образующие крупномасштабную структуру Вселенной, напоминающую паутину.

Скопления галактик представляют большой интерес для исследования, так как их свойства сильно зависят от космологических параметров. Изучая их свойства, можно делать выводы о структуре обозримой части Вселенной. По большей части скопления состоят из тёмной материи, природа которой до сих пор не известна науке. Поэтому через историю развития скоплений галактик можно изучать тёмную материю.

Сама по себе крупномасштабная структура Вселенной имеет объяснение. Через какое-то время после появления Вселенной возмущения волн плотности средних и больших масштабов при совпадении пиков образовали сверхскопления, в то время как сопадения фаз низкой плотности образовали войды - огромные пространства между нитями скоплений, в которых почти отсутствуют галактики и скопления. Таким образом, зная расположение и параметры большого количества скоплений, можно сделать выводы о том, как развивалась Вселенная на поздних этапах. [1]

Скопления демонстрируют сигналы на небе по всему электромагнитному спектру, что даёт три основных способа наблюдательного обнаружения этих систем: в виде избыточной плотности галактик в оптических и ближних инфракрасных обзорах, в качестве источников расширенного внегалактического излучения в рентгеновских длинах волн, а также по их сигнатуре Сюняева-Зельдовича (СЗ) в микроволновых обзорах. Каталоги, составленные по разным диапазонам, могут дополнять друг друга.

2.1 Оптический диапазон

Самым первым из каталогов скоплений стал каталог Abell [2], впервые опубликованный в 1958 г. Этот каталог содержит 4073 богатых

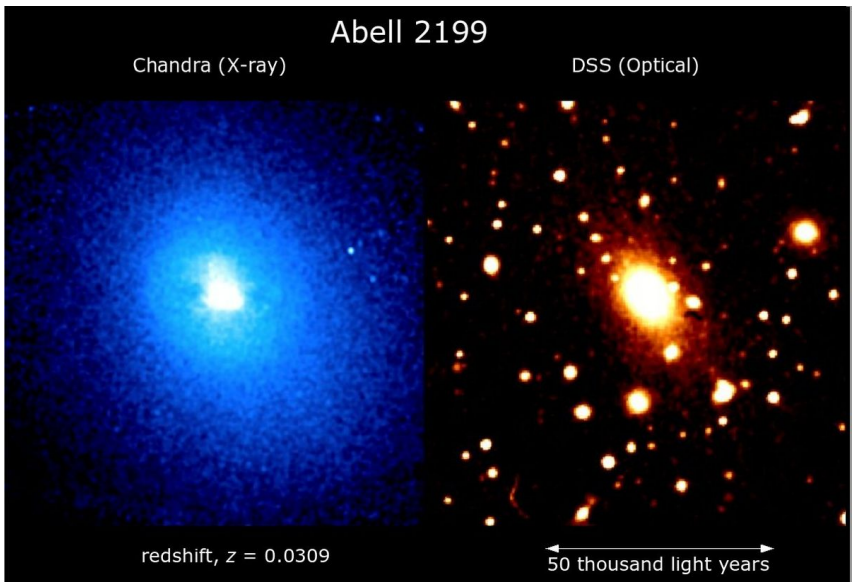

Рисунок 2.1 — Скопление Abell 2199 в рентгеновском и оптическом диапазонах

скопления галактик с красными смещениями $z < 0.2$. Он был построен при помощи ручного анализа оптических данных обзора NGS-POSS.

Логическим продолжением этого каталога является redMaPPer - каталог, созданный на основе обзора SDSS с помощью одноимённого алгоритма, что состоит из двух фаз - калибровки параметров «красной последовательности» как функции красного смещения и фазы поиска скоплений, где полученная на предыдущей фазе модель используется для поиска скоплений и измерения их богатства.

2.2 Рентгеновский диапазон

MCXC - компиляция каталогов скоплений, все из которых основаны на данных разных рентгеновских обзоров.

ROSAT - ренгтгеновский каталог по данным ROSAT, самый глубокий и чистый каталог до появления eRosita.

eROSITA (extended ROentgen Survey with an Imaging Telescope Array) — это рентгеновский телескоп, построенный Институтом внеземной физики Общества Макса Планка (MPE) в Германии. Его можно рассматривать как развитие рентгеновского телескопа ROSAT на современном научном и технологическом уровне. Рентгеновский зеркальный телескоп eROSITA интегрирован в космическую обсерваторию «Спектр-РГ»

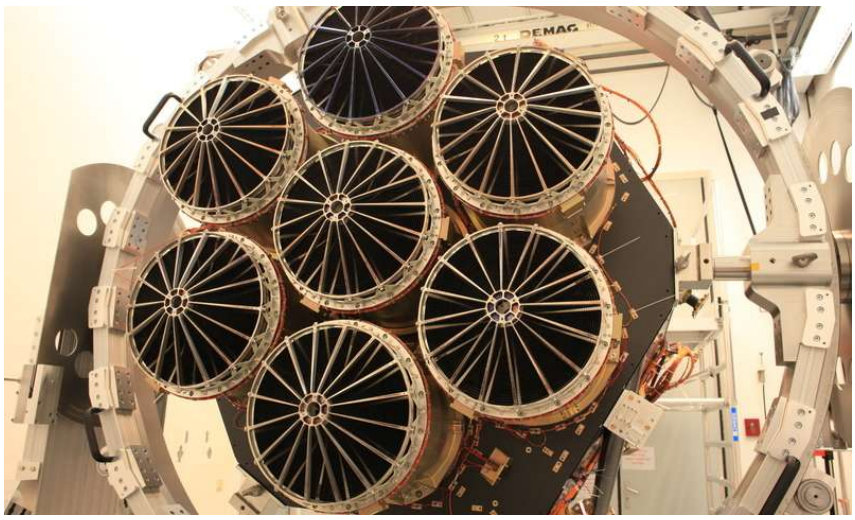

Рисунок 2.2 — Спутник eROSITA

(«Спектр-Рентген-Гамма»), которая включает также российский телескоп «ART-XC». Запуск обсерватории осуществлён компанией Роскосмос 13 июля 2019 года в 12:30:57 (UTC). Рентгеновский телескоп eROSITA снимет всё небо и составит восемь карт неба в мягком рентгеновском излучении. Ожидается, что рентгеновский телескоп eROSITA обнаружит 100 000 скоплений галактик. Основная научная цель состоит в том, чтобы измерить темную энергию через структуру и историю развития Вселенной, отслеживаемую скоплениями галактик.

eRosita - каталог, основанный на данных рентгеновского телескопа eRosita, основной целью которого и является поиск скоплений галактик.

2.3 Микроволновой диапазон

Эффект Сюняева - Зельдовича - изменение интенсивности радиоизлучения реликтового фона из-за обратного эффекта Комптона на горячих электронах межзвёздного и межгалактического газа. Эффект назван в честь предсказавших его в 1969 году учёных Р. А. Сюняева и Я. Б. Зельдовича.

С помощью эффекта Сюняева — Зельдовича можно измерить диаметр скопления галактик, благодаря чему скопления галактик могут быть использованы в качестве стандартной линейки при построении шкалы расстояний во Вселенной.

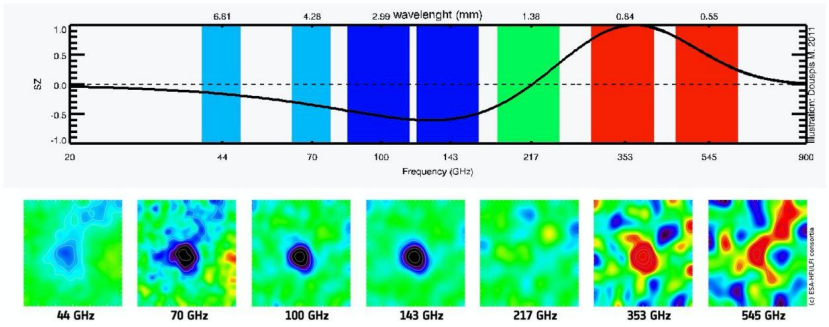

Fig. 2. Abell 2319 seen by *Planck* between 44 GHz and 545 GHz.

Рисунок 2.3 — Проявление скопления на картах Planck при помощи эффекта Сюняева-Зельдовича

Planck — астрономический спутник Европейского космического агентства (ЕКА), созданный для изучения вариаций космического микроволнового фона — реликтового излучения. Запущен 14 мая 2009 года ракетой-носителем «Ариан-5». В период с сентября 2009 по ноябрь 2010 года «Планк» успешно закончил основную часть своей исследовательской миссии, перейдя к дополнительной, завершившейся 23 октября 2013 года.

PSZ2 - каталог, основанный на данных микроволнового обзора Planck. При его создании использовались две версии алгоритма Matched Multi-filter, а также алгоритм PowellSnakes, основанный на байесовских методах.

Также существует каталог SPT, основанный на обзоре SPT-ECS.

ACT - каталог, основанный на данных микроволнового обзора ACT. Был создан при помощи алгоритма Matched Multi-Filter. На данный момент является самым полныым каталогом по микроволновому диапазону, однако он ограничен областью, на которой находятся данные ACT.

2.4 Актуальность

Создание каталогов данных в других диапазонах может помочь уточнить данные рентгеновских каталогов, созданных классическими алгоритмами.

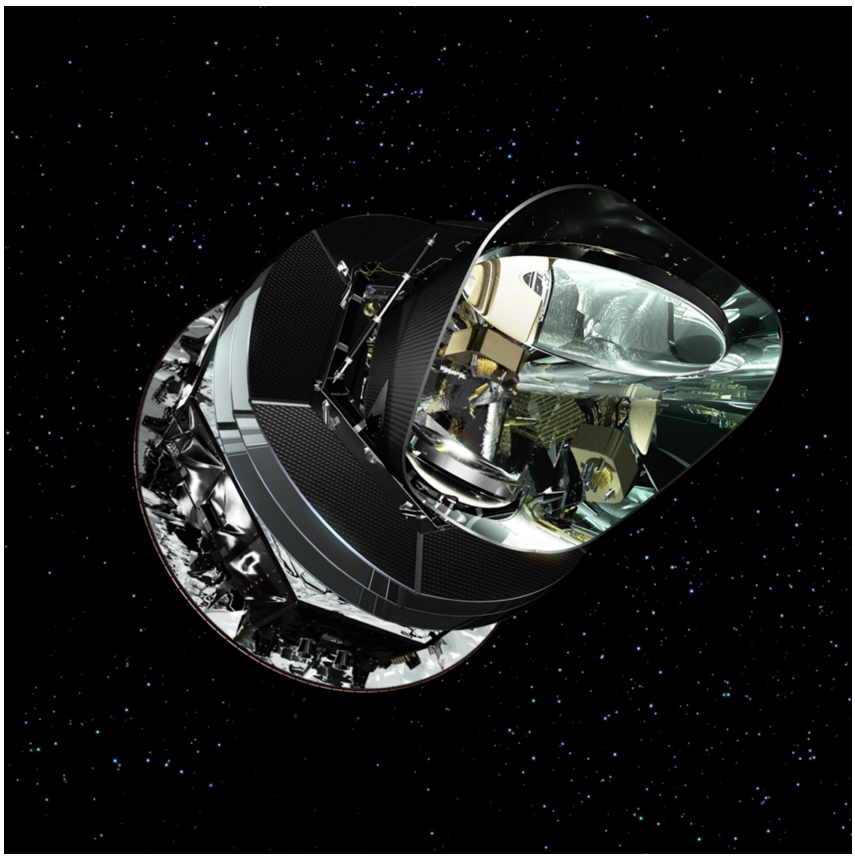

Рисунок 2.4 — Спутник Planck

Кроме того, сама задача применения нейросетевых методов к детекции скоплений актуальна, так как методы глубокого обучения дают следующие преимущества при анализе данных:

— Стандартные алгоритмы при детекции не используют всю доступную информацию, и данные всегда оказываются сложнее, чем математическая модель, призванная детектировать скопления; в то время как с помощью нейросети можно охватить данные полностью.

— Каждый из классических методов для каждого диапазона излучения не может быть совмещен с другим алгоритмом для другого диапазона, в то время как нейросеть может стать средством для сегментации данных нескольких каналов одновременно.

Методы глубокого обучения оказались крайне полезными во многих прикладных областях. С их помощью можно решать задачи распознавания образов, кластеризации, управления, прогнозирования, анализа данных, оптимизации, а также многие другие проблемы.

3 Постановка задачи

Исследование и разработка нейросетевых методов сегментации, детекции и классификации скоплений галактик в многоволновых данных обзоров неба, а также построение каталогов скоплений.

4 Обзор существующих решений

4.1 Методы детекции скоплений в данных Planck

PSZ2 - каталог, составленный по данным карт Planck с помощью трех методов: две версии Matched Multi-Filter и PowellSnakes.

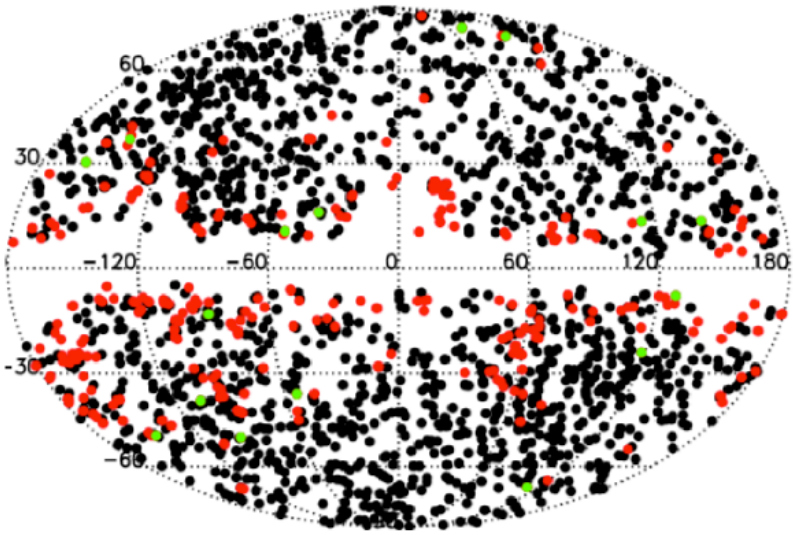

Рисунок 4.1 — Распределение необработанных обнаружений: удаленные кандидаты, помеченные инфракрасным флагом, отображаются красным цветом, а сохраненные обнаружения, помеченные инфракрасным флагом, - зеленым. Черные точки обозначают обнаружения без ИК-флага.

4.1.1 Метод согласованного мультифильтра: MMF1 и MMF3

Matched Multi-Filter, полученный с помощью моделирования, [3] позволяет восстановить оценку величины комптонизации излучения в картах Planck. MMF разделяет карты всего неба Planck на тангенциальные карты размером 14.66/10 (для MMF1/MMF3). После этого каждая карта фильтруется MMF, а затем регистрируются пики сигнала, превышающие заданный порог отношения сигнал-шум (заданный порог равен 4). Эти пики становятся координатами кандидатов в скопления.

Для MMF3 нужно ещё раз просканировать карты Planck, но в областях, центрированных на координатах, полученных на предыдущем

шаге. Фильтр применяется снова, что позволяет получить более точные значения размера и потока.

4.1.2 PowellSnakes (PwS)

PowellSnakes (PwS)[4] - это быстрый, полностью байесовский, многочастотный алгоритм обнаружения, разработанный для идентификации и определения характеристик компактных объектов, скрытых в диффузном фоне. PwS использует около 2800 квадратных участков со стороной 14,66 градуса, чтобы обеспечить надежное покрытие неба.

PwS обнаруживает скопления-кандидаты и в то же время вычисляет коэффициент надёжности и производит выборку из апостериорных распределений параметров скопления. Затем он объединяет подкаталоги из каждой карты-участка и применяет критерии для принятия или отклонения обнаружения.

Для обнаружения скоплений авторы применяют плоские априорные значения для положения и неинформативные априорные значения для радиуса и интегрированного потока, как определено с помощью метода Джеффри.

PwS может произвести небольшое количество ложных детекций с большим значением сигнал-шум, связанных с излучением галактической пыли. Авторы применяют дополнительное отсечение объектов, детектированных только PwS при $S/N > 10$.

Далее для объединения трёх подкаталогов процедура объединения идентифицирует обнаружение с наивысшим отношением сигнал-шум как опорную позицию во время слияния, а любые обнаружения из других подкаталогов в пределах 5 идентифицируются как опорная позиция.

Как отмечают авторы каталога PSZ2, полнота (Монте-Карло) систематически ниже, чем аналитическое приближение для полного обзора. Одна из причин этого - загрязнение галактической пылью, которое сильнее на более чем 20% неба, включенного в полную область обзора, по сравнению с областью космологической выборки. Это имеет тенденцию к снижению отношения сигнал / шум скоплений на некоторых линиях видимости.

4.1.3 Нейросетевые подходы

В статье [5] решается задача детекции скоплений в данных Planck. Авторы используют архитектуру U-net и каталоги PSZ2, MCXC и redMaPPer для обучения нейросетевой модели. Для сравнения с результатами этой статьи мы создали модель с такими же характеристиками, обучали на каталогах выбранных таким же образом и выбрали такие же области для тренировки и валидации.

4.2 Детекция скоплений по эффекту Сюняева-Зельдовича в данных SPT

Авторы статьи [6] моделируют с помощью гауссианы инструментальный и остаточный атмосферный шумы, чтобы создать фильтр, позволяющий на разных масштабах выявить кандидатов в скопления.

4.3 Детекция скоплений по эффекту Сюняева-Зельдовича в данных ACT

[7] Авторы каталога ACT моделируют сигнал скопления, используя Universal Pressure Profile, который подтверждается соответствующим лучом ACT для каждой частоты, чтобы сформировать шаблон сигнала S. Чтобы повысить эффективность обнаружения скоплений с разными угловыми размерами, они создали набор из 16 согласованных фильтров.

5 Исследование и решение задачи

5.1 Обзор данных

Основные данные для поиска скоплений - HFI-частотные карты Planck. Для создания тренировочной выборки использовались два каталога: PSZ2 и ACT. PSZ2 считается каталогом самых надежных скоплений, в то время как ACT считается самым полным каталогом (но для ограниченной области неба).

5.1.1 Данные Planck

Астрономический спутник Planck оснащен внеосевым телескопом системы Грегори. Зеркало фокусирует собранное излучение на два прибора: низкочастотный приёмник (LFI) (30-70 ГГц) и высокочастотный приёмник (HFI) (100-857 ГГц). Для детекции скоплений будут использоваться данные высокочастотного приёмника.

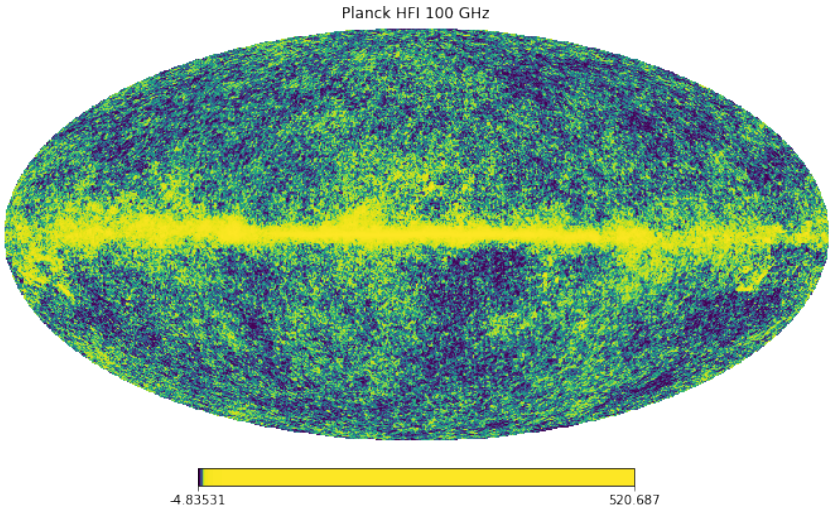

Рисунок 5.1 — Данные Planck на частоте 100 ГГц

Высокочастотный приёмник (HFI) — массив из 54 болометрических детекторов, которые преобразуют принятое излучения в тепло. Количество тепла затем измеряется электрическим термометром, сигнал с которого преобразуется в температуру с помощью компьютера. HFI детекторы работают в шести частотных каналах в интервале от 100 до 857 ГГц. Они охлаждены практически до абсолютного нуля, работая при температуре $-273C$.

5.1.2 Каталоги скоплений по данным микроволнового диапазона

5.1.2.1 Каталог PSZ2

[1] Это каталог всего неба источников Сюняева-Зельдовича (SZ), обнаруженных по полным 29-месячным данным миссии Planck. Каталог (PSZ2) - это самая большая выборка скоплений галактик, отобранная по SZ, и самый глубокий систематический обзор скоплений галактик по всему небу. Он содержит 1653 обнаружения, из которых 1203 являются подтвержденными скоплениями с идентифицированными аналогами во внешних наборах данных. В справочной статье авторы описывают многоволновой поиск аналогов во вспомогательных данных, который использует наборы радио-, микроволновых, инфракрасных, оптических и рентгеновских данных и делают упор на надежность сопоставления двойников. Они обсуждают физические свойства нового каталога и идентифицируют совокупность тусклых рентгеновских скоплений с малым красным смещением, выявленных с помощью SZ-отбора. Эти объекты появляются в оптических обзорах и обзорах SZ с одинаковыми характеристиками для их массы, но они почти отсутствуют в отобранных рентгеновских выборках ROSAT.

Для обнаружения скоплений SZ использовались три метода: две независимые реализации согласованного мультифильтра (MMF1 и MMF3) и PowellSnakes (PwS). Главный каталог построен как объединение каталогов трех методов.

5.1.2.2 Каталог ACT

[7] Это каталог из 4195 оптически подтвержденных скоплений галактик Сюняева-Зельдовича (SZ), обнаруженных на $13168 deg^2$ неба (примерно 32% всего неба), обследованных Космологическим телескопом Атакама (ACT). Кандидаты в кластеры были отобраны путем применения многочастотного согласованного фильтра к картам 98 и 150 ГГц, построенным на основе всех наблюдений ACT, полученных в 2008–2018 гг., и впоследствии подтвержденных с помощью глубоких оптических обзоров

с большой площадью. Обнаруженные кластеры охватывают диапазон красного смещения $0{,}04 < z < 1{,}91$ со средним значением $z = 0{,}52$. Каталог содержит 221 кластер с $z > 1$, а всего 872 системы являются новыми открытиями. Выборка скоплений более чем в 22 раза больше, чем предыдущий каталог скоплений ACT, и на сегодняшний день является самой большой однородной выборкой скоплений, выбранных SZ. Зона обзора имеет большое перекрытие с глубокими оптическими исследованиями со слабым линзированием, которые используются для калибровки отношения масштабирования массы SZ-сигнала, такими как исследование темной энергии (Dark Energy Survey) ($4552 deg^2$), стратегическая программа Hyper Suprime-Cam Subaru ($468 deg^2$) и Kilo Degree Survey ($823 deg^2$).

5.1.3 Каталоги скоплений по данным оптического диапазона: redMaPPer

Каталог[8] из около 25000 скоплений в диапазоне красных смещений $z \in [0{,}08, 0{,}55]$. Фотометрические красные смещения redMaPPer близки к гауссовым, с разбросом $z \approx 0{,}006$ при $z \approx 0{,}1$, возрастающим до $z \approx 0{,}02$ при $z \approx 0{,}5$ из-за увеличения фотометрического шума вблизи предела обзора. Среднее значение $|\Delta z|/(1 + z)$ для полной выборки составляет 0,006. Частота возникновения проекционных эффектов низкая ($\leqslant 5\%$). Для оценки M500 скоплений нужно использовать зависимость от параметра богатства и красного смещения. [9]

5.1.4 Рентгеновкие каталоги скоплений

5.1.4.1 MCXC

MCXC - это мета-каталог собранных свойств скоплений галактик, обнаруженных в рентгеновских данных. Этот очень большой каталог основан на общедоступных скоплениях ROSAT All Sky Survey (RASS) (NORAS, REFLEX, BCS, SGP, NEP, MACS и CIZA) и на кластерах ROSAT (160SD, 400SD, SHARC, WARPS и EMSS). Данные были тщательно обработаны для удаления повторяющихся записей из-за перекрытия между областями исследования отдельных входных каталогов. MCXC со-

стоит из 1743 скоплений практически без повторяющихся записей. Для каждого скопления MCXC предоставляет красное смещение, координаты, принадлежность к исходному каталогу, общую массу M500 и радиус R500. Мета-каталог дополнительно предоставляет информацию о перекрытиях между входными каталогами и отношениями светимости, гдк доступны измерения из разных обзоров, и дает примечания по отдельным объектам. MCXC доступен для обеспечения максимальной полезности для рентгеновских исследований, исследований эффекта Сюняева-Зельдовича (SZ) и других многоволновых исследований.

5.1.4.2 2RXS

2RXS [10] - ренгтгеновский каталог по данным ROSAT, самый глубокий и чистый каталог по рентгеновским данным до появления каталога eRosita.

5.2 U-net

U-net [11] 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. Для данной работы архитектура была реализована с помощью библиотеки tensorflow.keras.

5.3 Другие нейросетевые модели

На данный момент для решения задачи сегментации карт Planck использовалась только модель U-net, однако кроме неё рассматривались и другие варианты.

5.3.1 PSPNet

[12] Эта модель была создана для использования в области scene parsing. Главная задача, которую эта модель помогает решить - попик-

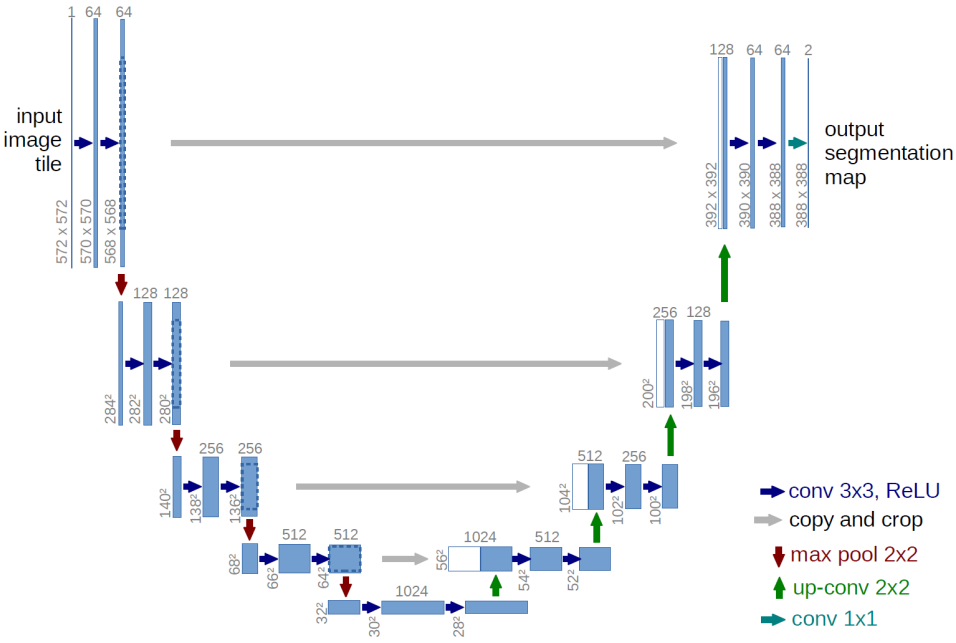

Рисунок 5.2 — Структура модели U-net

сельная сегментация объектов на изображении при условии наличия большого количества меток (например, датасет ADE20K, о котором идёт речь в статье, содержит изображения с 150 метками).

PSPNet помогает решить проблемы:

— взаимосвязи меток (например, объект с меткой «самолёт» скорее всего будет находиться в пространстве с меткой «аэропорт» или «посадочная полоса»)

— совпадающих категорий (наличие в тренировочной выборке объектов вроде «поле» и «земля», «холм» и «гора»)

— небольших объектов (например, «фонарь» или «вывеска», находящиеся на дальнем плане изображения)

Из вышеперечисленных проблем в текущей работе по сегментации скоплений нас может интересовать только последняя - искомое скопление может занимать маленькую площадь относительно всего изображения.

Основная идея PSPNet заключается в использовании Pyramid Pooling Module. Чтобы получить этот модуль, нужно сначала сделать несколько версий изначального изображения в разных масштабах с помощью pooling слоёв разных размеров, первый «грубый» слой собирает данные всего изображения в один многоканальный пиксель, следующий

содержит данные нескольких подрегионов, и далее каждый последующий содержат всё меньше глобальной информации и всё больше локальной.

Далее для каждого масштаба проводится свёртка для уменьшения количества каналов и upscaling с помощью билинейной интерполяции, и все масштабы конкатенируются, чтобы с помощью последнего слоя свёртки получить для них итоговую сегментацию.

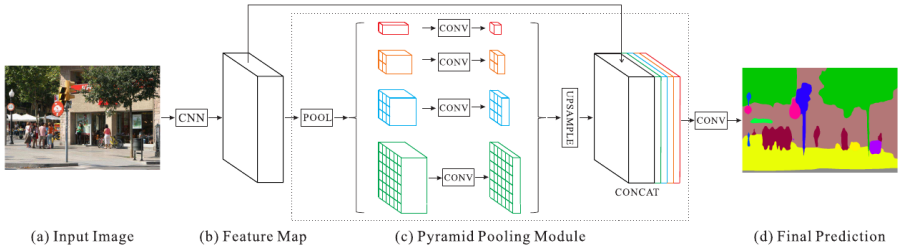

Рисунок 5.3 — Структура модели PSPNet

5.3.2 LinkNet

[13] Эта модель решает проблему большого количества параметров и низкой скорости других моделей из похожей области применения. LinkNet хорошо подойдёт для сегментации в реальном времени и сегментации на видео, что тоже имеет очень слабое отношение к нашей области, где количество данных фиксировано и сегментация в реальном времени не требуется, однако скорость нейросетевой модели тем не менее является плюсом.

Общая структура LinkNet очень напоминает UNet, с тем отличием, что в блоках кодировщика добавлены дополнительные skip-connection связи.

5.3.3 W-Net

[14]Эта архитектура была призвана улучшить результаты сегментации при условиях:

— без использования сложных архитектур свёрточных нейронных сетей

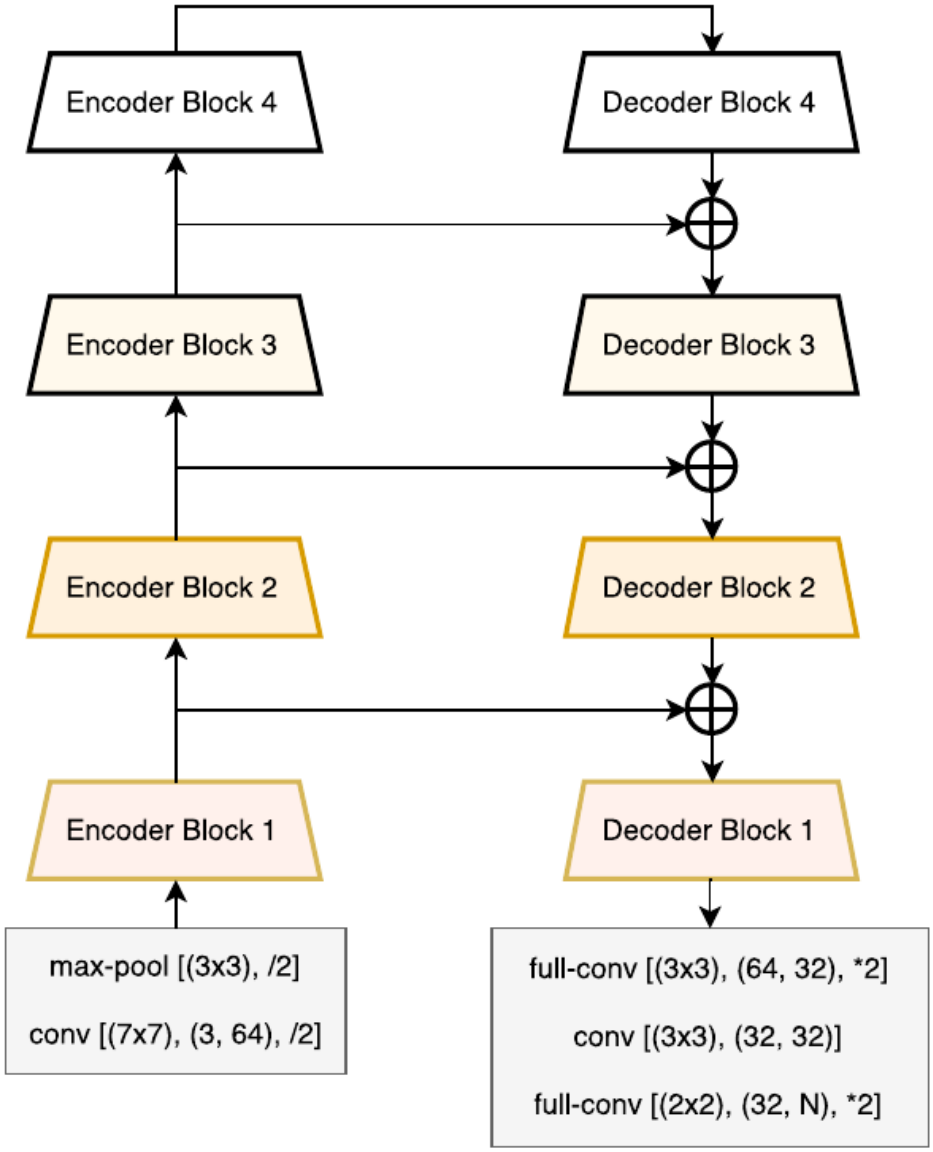

Рисунок 5.4 — Структура модели LinkNet

— при тестировании на данных из разных датасетов

При сегментации космических данных мы обычно имеем общий обзор некоторой области неба в определенном спектре, и эти данные получены с помощью определенного телескопа. Однако возможно появится причина обучать модель на данных одного телескопа, а тестировать на других, что в какой-то степени кореллирует с описанными условиями.

Архитектура W-Net предлагает улучшение модели U-Net: обучение происходит на двух последовательных нейросетевых моделях. Выход первой модели конкатенируется с входом и отправляется на

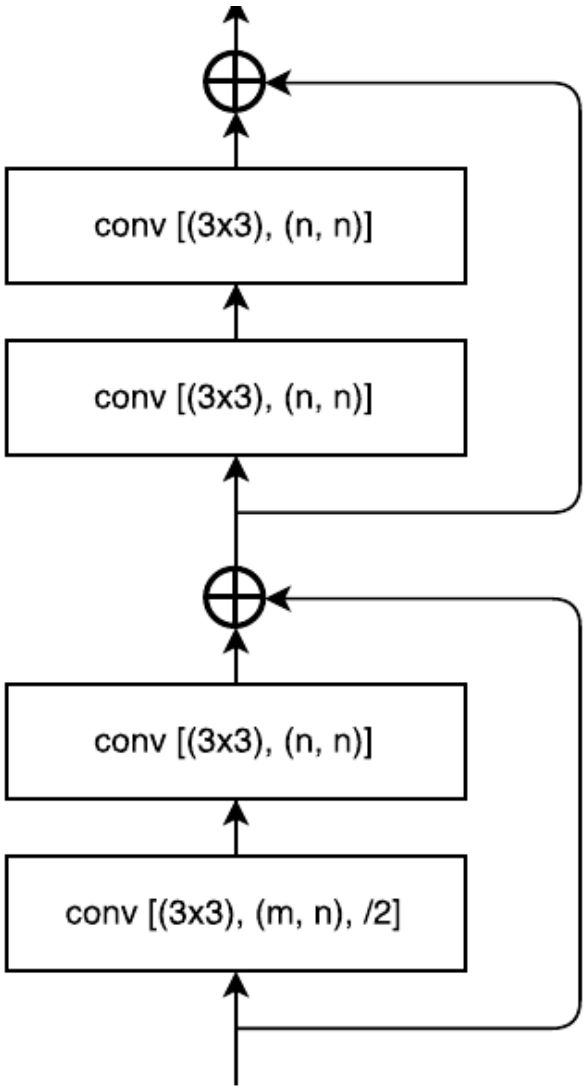

Рисунок 5.5 — Кодировщик модели LinkNet

обработку второй моделью.

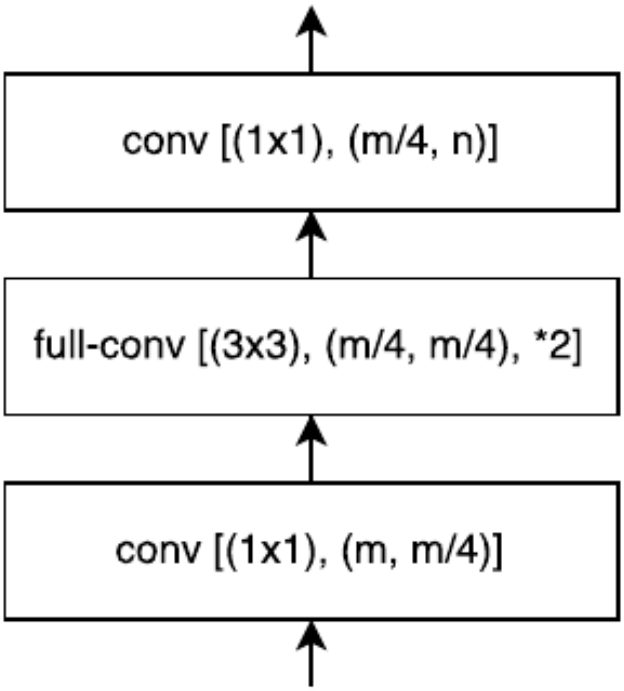

Рисунок 5.6 — Декодировщик модели LinkNet

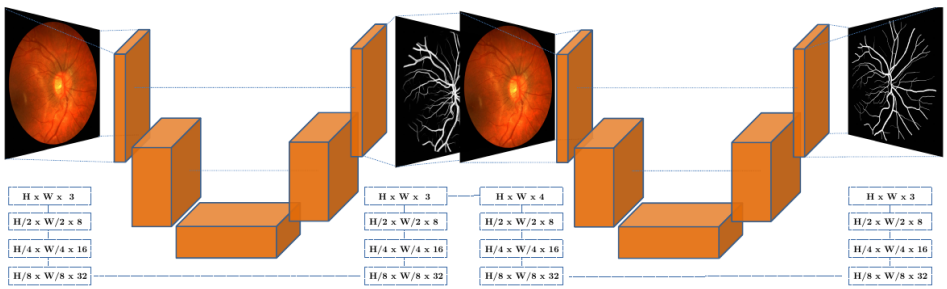

Рисунок 5.7 — Структура модели WNet

6 Практическая часть

6.1 Модель

Модель в статье имеет такую структуру: U-net с 5-ю блоками в кодировщике, где каждый блок кодировщика содержит две свёртки 3×3 с функцией активации ReLU и слоём max-pooling (с увеличением количества фильтров в 2 раза). Блок кодировщика состоит из 2×2 upsampling свёртки (с уменьшением количества фильтров в 2 раза), конкатенации выхода соответствующего блока из кодировщика, двух свёрток 3×3 с ReLU. Самый последний слой кодировщика дополняется свёрткой 1×1, кроме того, после каждой свёртки добавлен слой Dropout с параметром 0.2, чтобы увеличить эффективность обучения. Для изменения весов нейросети использовался оптимизатор Adam с параметром $lr = 10^{-4}$. В качестве loss-функции выбрана бинарная кросс-энтропия. Размер батча 20, количество эпох варьировалось от 20 до 40.

6.2 Предобработка данных и обучение модели

6.2.1 Нормализация данных

В первую очередь, нужно нормализовать данные Planck, подбирая для каждого канала параметры гауссианы, которая покрывает большую часть нужных данных. Таким образом сохранено изначальное распределение, но значения были привдены к формату, который будет лучше восприниматься нейросетевой моделью.

6.2.2 Каталоги для обучения

Для обучения нейросетевой модели мы использовали следующие каталоги:

— PSZ2
— MCXC
— RedMaPPer

Некоторые из этих каталогов содержат общие объекты, кроме того, обучать модель сразу на всех данных может оказаться нецелесообраз-

но (не все объекты из этих каталогов возможно найти в данных Planck), поэтому списки были переработаны следующим образом:

— planck_z - каталог скоплений PSZ2 с известным красным смещением

— planck_no_z - каталог скоплений PSZ2 без красных смещений

— mcxcwp - каталог тех скоплений MCXC, что не присутствуют в PSZ2

— RM30 и RM50 - каталоги RedMaPPer

6.2.3 Разбиение данных на выборки

После этого вся область неба разделяется на три части - тренировочная, валидационная и тестовая. Для такого разбиения используется алгоритм представления данных на сфере HEALPix с параметром $n_{side} = 2$. В качестве валидационных выбраны пиксели $9, 38, 41$, в качестве тестовых - 6, все остальные 44 пикселя неба использовались для создания данных для тренировочной выборки (в статье [5] эти пиксели нумеруются с 1, здесь с 0).

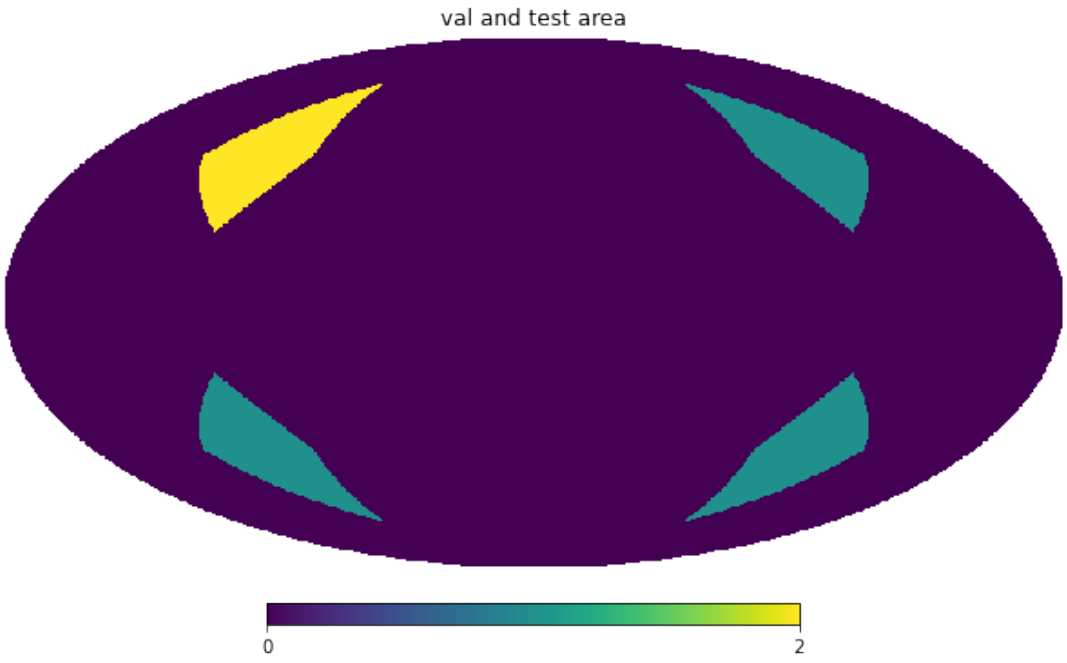

Рисунок 6.1 — Валидационная и тренировочная области

6.2.4 Генерация патчей и масок

Для этой работы будет использоваться нейросетевая архитектура U-net для сегментации данных Planck и создания своего собственного каталога детектированных объектов. Для обучения будем использовать данные Planck в виде квадратных патчей размером 1.83×1.83, (для 6 HFI-каналов форма матрицы патча составит $64 \times 64 \times 6$), а также дополним их масками, обозначающими расположение скоплений. На каждой маске единицами отмечены пиксели в радиусе $2{,}5'$ от скопления из тренировочного каталога, а нулями помечены все остальные пиксели. Патчи для обучения выбираются так, чтобы в их области содержался хотя бы один центр скопления из тренировочного каталога.

В качестве тренировочного каталога использовались каталоги planck_z и planck_z + act (часть каталога act была отброшена).

6.2.5 Обучение модели

Модель обучалась до до 100 эпох. На каждой эпохе использоваось 100000 патчей, размер батча - 20. Для валидации использовалось 12000 патчей.

6.2.6 Метрики

Кросс-энтропия - классическая loss-функция для задачи сегментации.

$$K_{log} = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(p(y_i)) + (1 - y_i) * log(1 - p(y_i)$$

Recall - также одна из классических метрик для многих задач глубокого обучения.

$$Recall = \frac{|true\_positives|}{|true\_positives| + |false\_negatives|}$$

Кроме loss-функции бинарной кроссэнтропии для отслеживания результатов IoU и Dice.

IoU - это площадь области перекрытия между прогнозируемой сегментацией и истинной сегментацией, разделенной на площадь их объединения.

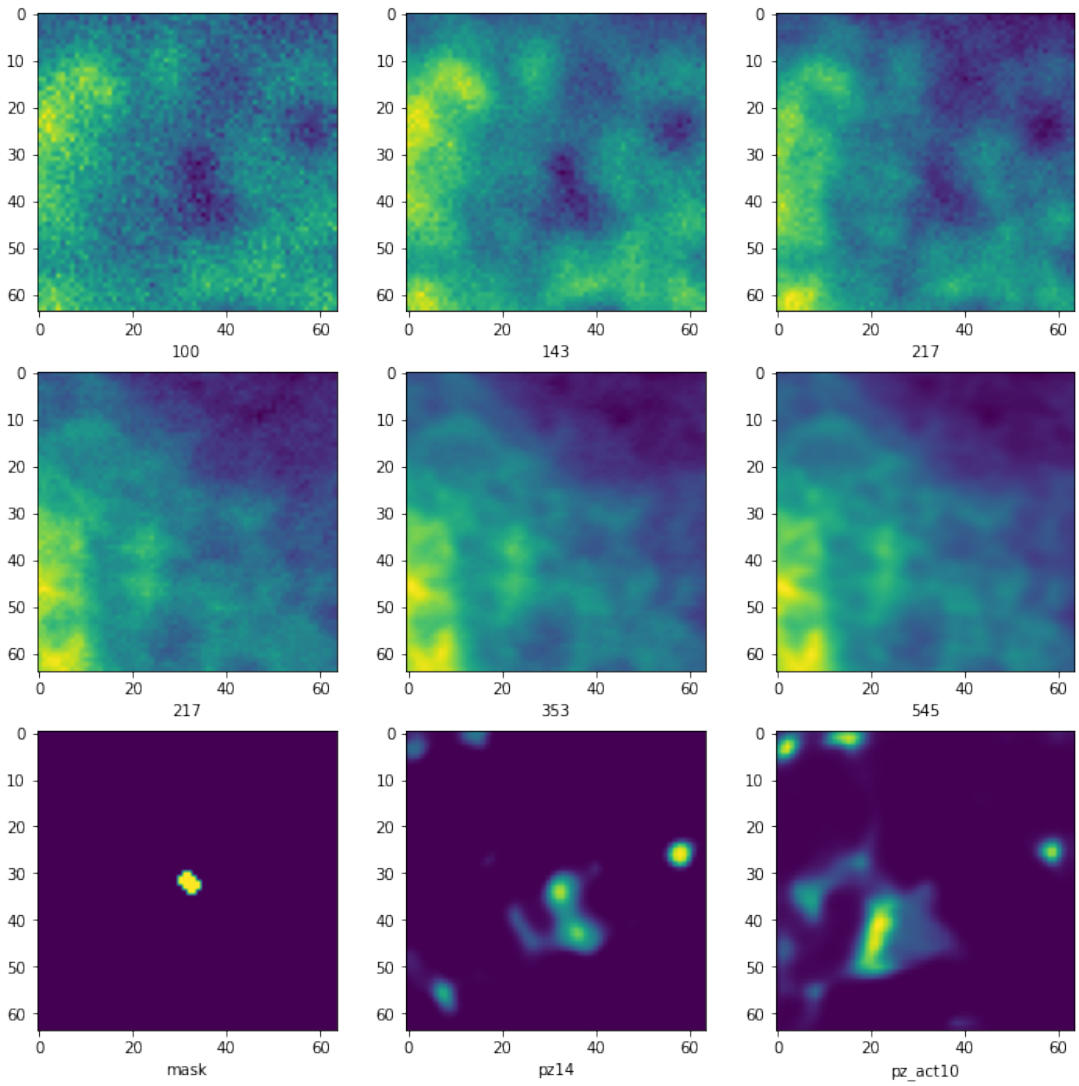

Рисунок 6.2 — Пример патча с маской скопления и предсказанынми масками для разных моделей

$$K_{IoU} = \frac{|A \cap B|}{|A \cup B|}$$

Коэффициент Dice - это удвоенная площадь перекрытия прогнозируемой сегментации с истинной сегментацией, деленная сумму их площадей.

$$K_{dice} = \frac{2*|A \cap B|}{|A| + |B|}$$

Метрика accuracy в целом не несёт основной информации для сегментации, так как в большинстве случаев предсказанная маска, состоящая полностью из нулей, будет иметь accuracy ≈ 0.9.

Выбранная loss-функция бинарная кроссэнтропия является классической метрикой для обучения моделей для сегментации. Её выбор

влияет не только на текущую задачу обнаружения кандидатов в скопления галактик, но и на другие задачи, которые можно изучать далее - например извлечение из карт Planck распределения тёмной материи.

Метрики сегментации IoU и Dice показывают похожие результаты и являются ещё одним способом отследить момент, когда модель начинает переобучаться.

Имея информацию обо всех этих метриках, можно выбрать подходящую модель для создания итогового каталога.

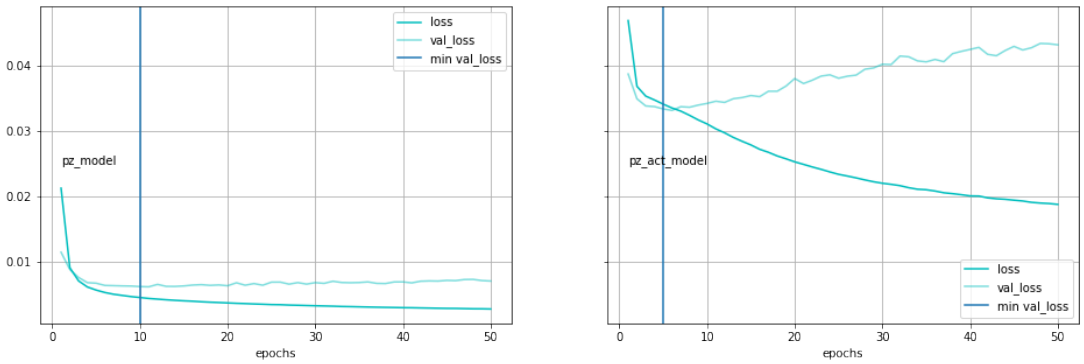

Рисунок 6.3 — График фукции loss для моделей, обученных на planck_z и planck_z + act по отношению к эпохе

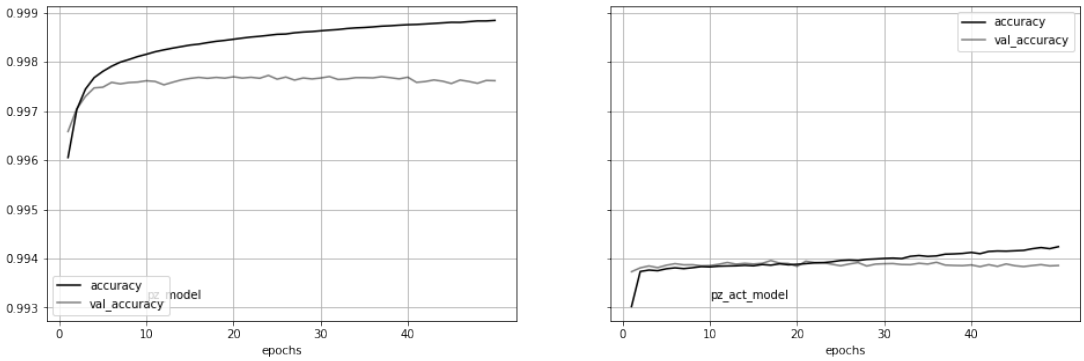

Рисунок 6.4 — График accuracy для моделей, обученных на planck_z и planck_z + act по отношению к эпохе

6.3 Детекция скоплений на карте Planck.

Остановимся подробно на этапе детекции:

а) Для сканирования выбирается один пиксель из разбиения HEALPix с $n_{side} = 2$ (всё небо разделяется на 48 таких пикселей);

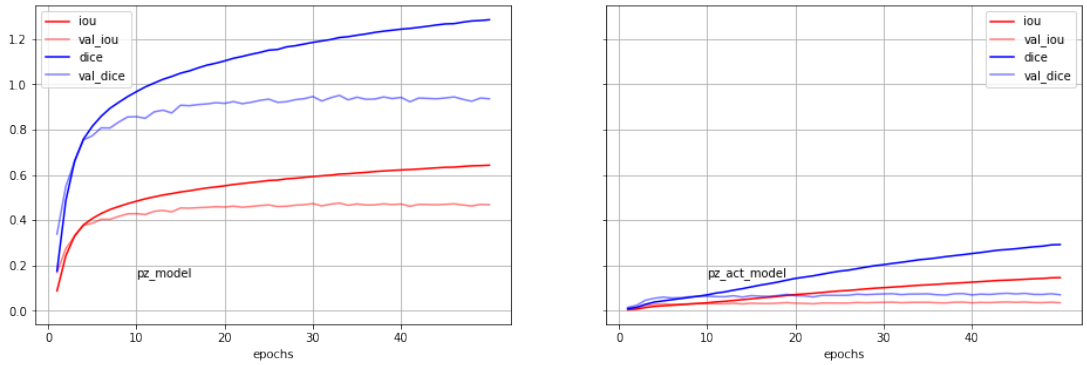

Рисунок 6.5 — График IoU и dice для моделей, обученных на planck_z и planck_z + act по отношению к эпохе

б) Чтобы просканировать всю область нейросетью, нужно пройтись по ней окном 64×64. Для этого нужно выбрать шаг сканирования - можно задать его как 64, тогда сканирование получится без пересечений, однако при меньшем шаге результаты оказываются лучше (если рассматривать полноту по выбранным каталогам и количество неизвестных объектов). Таким образом с помощью окна большой пиксель разбивается на патчи;

в) Каждый такой патч отправляется на вход в нейросеть. Полученные маски склеиваются и усредняются, чтобы снова получить изображение размером с пиксель $n_{side} = 2$;

г) На полученной маске отделяются области со значениями, превышающими заданный порог;

д) Для каждой области определяется её барицентр, который преобразуется в небесные координаты;

Все остальные детали, такие как архитектура нейросетевой модели, выбранные области для тренировочной, валидационной и тестовой выборки, размер тренировочной выборки повторяют то, что было описано в [5].

На данный момент для детекции был выбран шаг сканирования 8, так как это наименьший шаг сканирования, при использовании которого тратится вменяемое количество времени, кроме того, из всех возможных вариантов этого параметра это значение дает наилучшие результаты recall и fp на тестовой области:

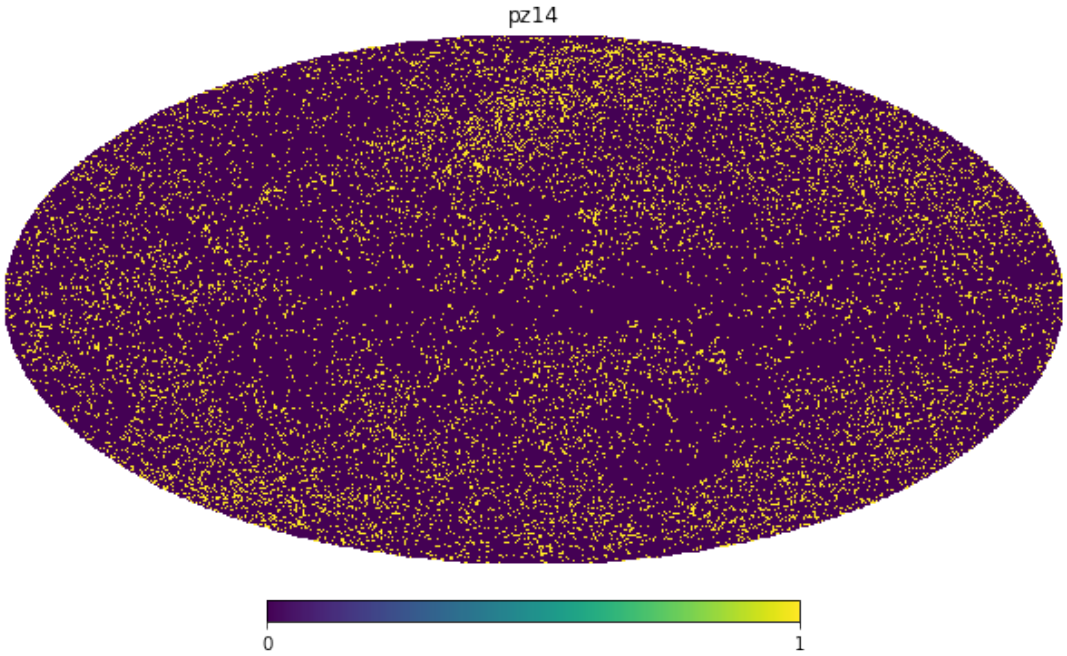

Рисунок 6.6 — Распределение объектов из детектированного каталога pz14 (каталог, полученный обучением на planck_z, отсканированный на 14 эпохе модели)

Однако предполагается, что при значения шага меньше 8 результаты могут быть еще лучше, но этот вопрос еще предстоит исследовать. С меньшим шагом сканирования слишком сильно увеличивается время вычисления детекции.

6.4 Формирование каталога скоплений

У полученных после детекции каталогов определены следующие параметры для каждого из детектированных объектов:

— area – площадь сегментированной области скопления;

— min_rad, max_rad, mean_rad – минимальный, максимальный, средний радиусы области;

— min_pred, max_pred – минимальное, максимальное значение маски в области;

— status – факт сопоставления скопления с объектом из каталога;

— catalog — сопоставленный каталог;

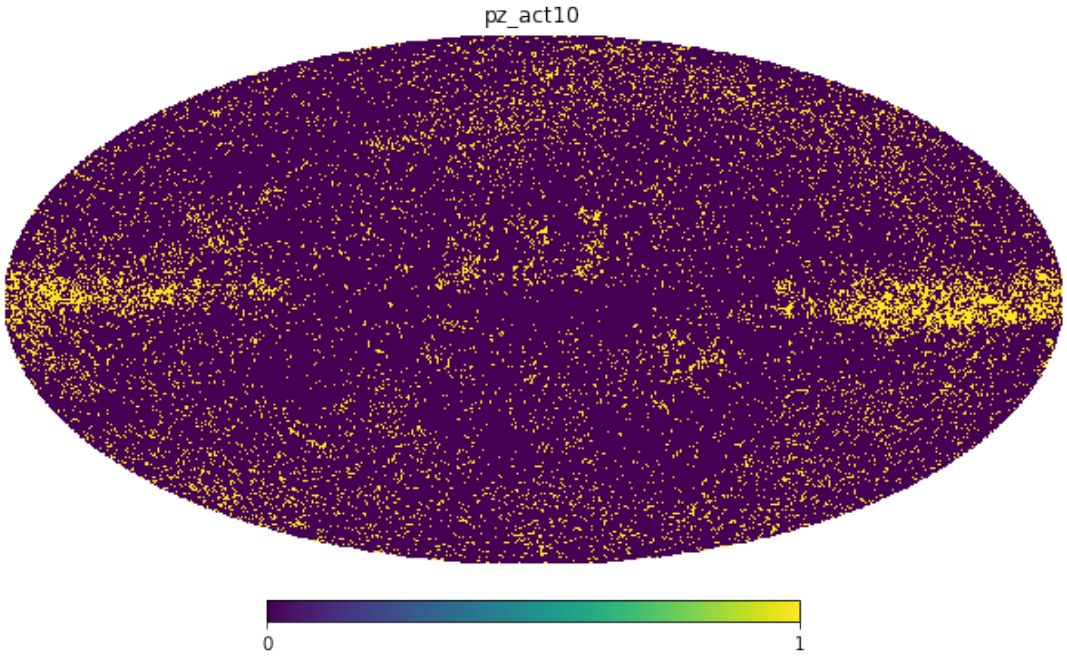

Рисунок 6.7 — Распределение объектов из детектированного каталога pz_act10 (каталог, полученный обучением на planck_z + act, отсканированный на 10 эпохе модели)

6.5 Выбор эпохи для сканирования

Как уже было упомянуто, среди 48 больших областей на небе три были выбраны как валидационные. Они содержат достаточное количество скоплений из каталога ACT, чтобы по ним можно было оценивать обе модели.

Количество false positive объектов меняется слишком часто, чтобы можно было выбирать модель по этому признаку, в то время как recall с определённого момента выравнивается по большей части каталогов. Начиная с 20-25 эпох это можно наблюдать, сравнивая recall для каталога ACT. В это же время recall для всего неба для ACT будет заметно выше для модели pz_act, что говорит о том, что на этих эпохах модель уже начинает переобучаться.

Рисунок 6.8 — Пример патча с шестью каналами данных Planck и маской с отмеченным центром скопления

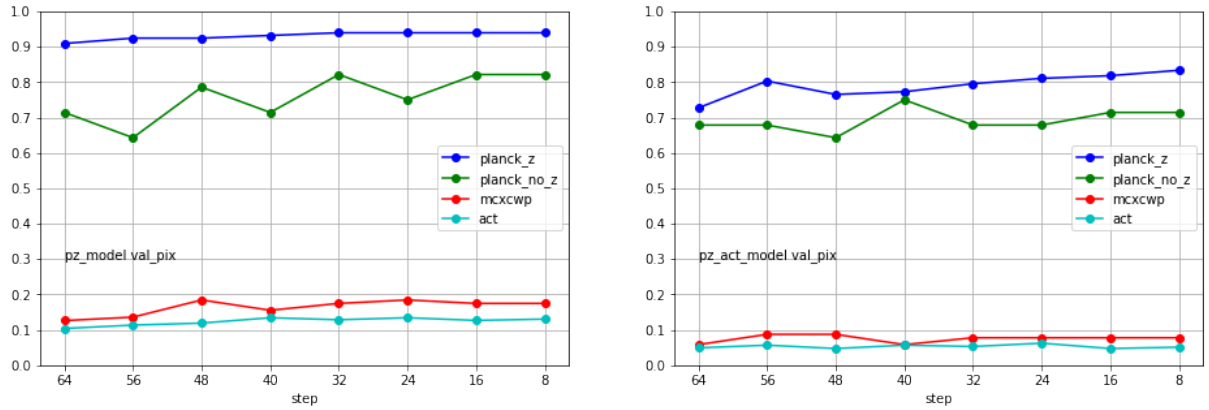

Рисунок 6.9 — Сравнение различных вариантов параметра шага сканирования

	PSZ2	MCXC	RM	ACT	fp
pz14	0.901996	0.415950	0.047566	0.202384	15828
pz20	0.924380	0.424555	0.050017	0.208582	23104
pz25	0.922565	0.428571	0.052622	0.211681	20611
pz40	0.911071	0.423982	0.048983	0.207628	17306
pz_act10	0.823351	0.378084	0.039409	0.220739	16316
pz_act14	0.824561	0.384395	0.046379	0.305602	16484
pz_act20	0.736237	0.343660	0.045268	0.444815	9398
pz_act25	0.762250	0.349971	0.048677	0.512992	15275

Таблица 6.1 — Recall по необрезанным каталогам для всего неба

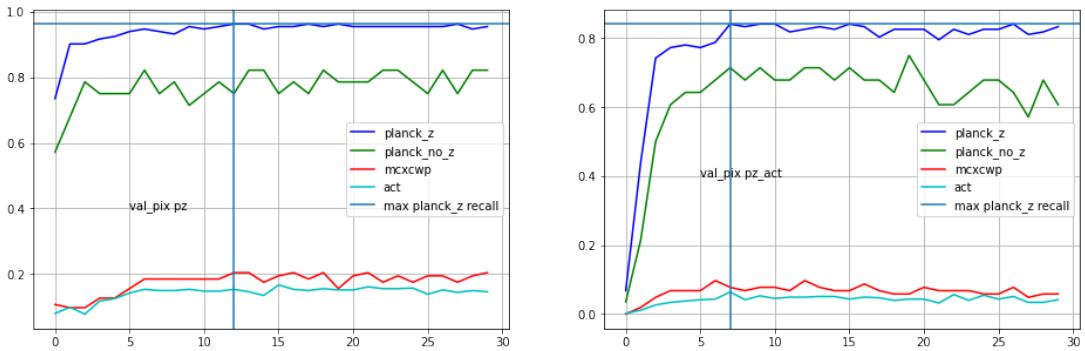

Рисунок 6.10 — Recall по каталогам planck_z, planck_no_z, mcxcwp, act для моделей pz и pz_act относительно эпохи на валидационной области

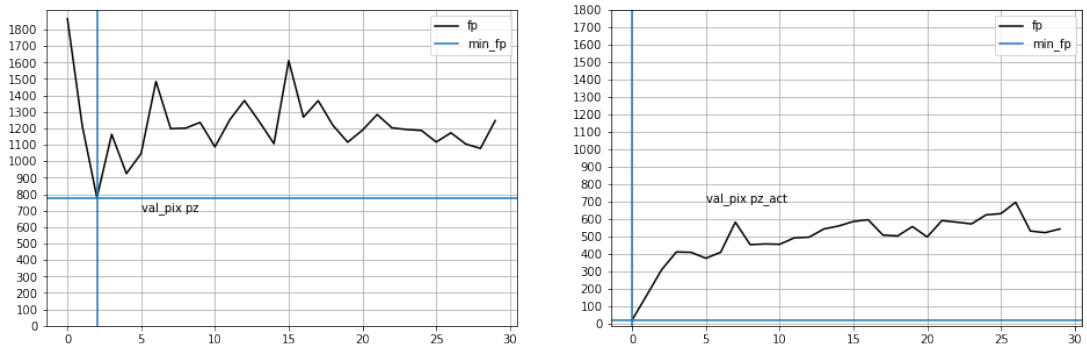

Рисунок 6.11 — Количество неопознанных объектов (false positive)для моделей pz и pz_act относительно эпохи на валидационной области

	PSZ2	MCXC	RM	ACT	fp
pz14	0.91875	0.427711	0.043817	0.210963	1055
pz20	0.93750	0.427711	0.046738	0.237542	1116
pz25	0.922565	0.428571	0.052622	0.211681	20611
pz40	0.91250	0.415663	0.046089	0.214286	1180
pz_act10	0.90000	0.385542	0.033755	0.157807	624
pz_act14	0.89375	0.373494	0.037001	0.157807	759
pz_act20	0.82500	0.349398	0.026939	0.119601	557
pz_act25	0.81875	0.349398	0.029860	0.127907	624

Таблица 6.2 — Recall по необрезанным каталогам для валидационной области

Заключение

Текущие результаты на данный момент:

— Созданы алгоритмы для предварительной обработки данных Planck.

— Обучены две модели с архитектурой U-net на HFI картах Planck и каталогах PSZ2 и PSZ2 + ACT.

— На базе этих моделей составлены каталоги кандидатов в скопления галактик. Эти каталоги сравнивались с каталогом eRosita и имеют с ним 1704 и 2355 совпадений соответственно.

Модель U-net обучалась на двух каталогах: planck_z и planck_z + act. Для каждой из этих моделей были выбраны две наиболее удачные по соотношению recall/количество false positives эпохи.

Для модели, обученной только на planck_z, такой эпохой стала 14, а для модели с planck_z + act это эпоха 10. Далее приведена таблица пересечений изначальных каталогов PSZ2, MCXC и ACT с итоговыми каталогами.

Каталог	PSZ2	err	MCXC	err	ACT	err
pz14	1491	$15.75^{\pm 1.07}$	725	$17.95^{\pm 0.84}$	849	$43.25^{\pm 1.39}$
pz_act10	1361	$13.20^{\pm 0.94}$	659	$13.65^{\pm 0.79}$	926	$30.65^{\pm 1.08}$
all	1653		1743		4195	

Таблица 6.1 — Сравнение детектированных каталогов с другими известными каталогами скоплений

Каталог	PSZ2	MCXC	ACT	fp
pz14	0.902	0.416	0.202	15828
pz_act10	0.823	0.378	0.22	16316

Таблица 6.2 — Отклик для детектированных каталогов по другим известным каталогам скоплений

Также были проведены сравнения с каталогами, для которых были выбраны объекты со значением $M500 > 4$. Отклик на этих каталогах оказался значительно выше, чем на полных каталогах, что говорит об особенности полученных моделей - они лучше подходят для обнаружения более массивных скоплений.

В дальнейшем, если получится выявлять оценки M500 для полученных объектов, то можно будет ранжировать каталоги по ним и выявлять наиболее надежных кандидатов в скопления.

Каталог	PSZ2	MCXC	ACT	fp
pz14	0.985	0.848	0.76	158288
pz_act10	0.926	0.83	0.728	16316

Таблица 6.3 — Отклик для детектированных каталогов по другим известным каталогам скоплений для скоплений со значением M500 > 4

Каталог	PSZ2	MCXC	ACT	fp
pz14	0.969	0.846	0.545	15828
pz_act10	0.876	0.846	0.515	16316

Таблица 6.4 — Отклик для детектированных каталогов по другим известным каталогам скоплений для скоплений со значением M500 > 4 и красным смещением $z > 0.5$

По показанным далее распределениям найденных скоплений можно заметить, что общее соотношение между z и M500 для моделей pz14 и pz_act10 почти не отличается. Лучше всего особенности каталогов проявляются на сопоставлении со скоплениями из каталога ACT: объекты разделяются на две группы, которые хоть и нельзя разделить, но можно увидеть в них значительное различие по диапазону M500.

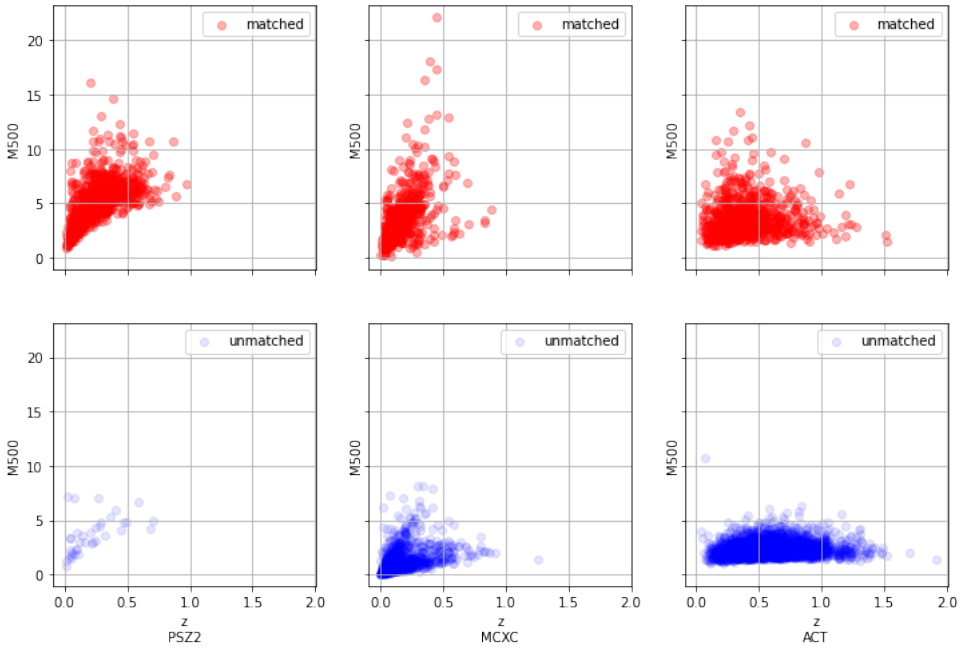

Рисунок 6.1 — Распределение найденных и не найденных скоплений из каталогов PSZ2, MCXC и ACT по z и M500 для каталога pz14

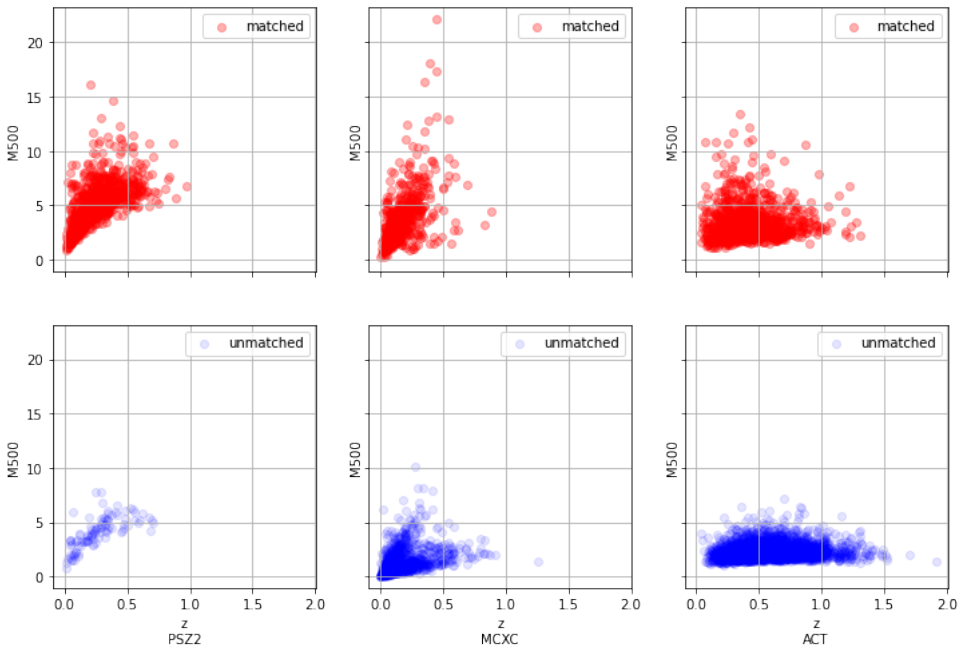

Рисунок 6.2 — Распределение найденных и не найденных скоплений из каталогов PSZ2, MCXC и ACT по z и M500 для каталога pz_act10

Список использованных источников

1. *collaboration, Planck.* Planck 2015 results. XXVII. The second Planck catalogue of Sunyaev-Zeldovich sources. / Planck collaboration. — Astron. Astrophys., 2015.

2. *Abell G. O., Corwin H. G. Jr.* A Catalog of Rich Clusters of Galaxies / Corwin H. G. Jr. Abell, G. O. — R. P. Astrophys. J. Suppl., 1989.

3. *Martin G. Haehnelt, Max Tegmark.* Using the kinematic Sunyaev-Zeldovich effect to determine the peculiar velocities of clusters of galaxies / Max Tegmark Martin G. Haehnelt. — 1995.

4. *Carvalho Pedro; Rocha, Graça; Hobson M. P.* A fast Bayesian approach to discrete object detection in astronomical data sets - PowellSnakes I / Graça; Hobson M. P. Carvalho, Pedro; Rocha. — 2009.

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

6. *collaboration, SPT.* GALAXY CLUSTERS DISCOVERED VIA THE SUNYAEV-ZEL'DOVICH EFFECT IN THE 2500-SQUARE-DEGREE SPT-SZ SURVEY / SPT collaboration. — 2015.

7. *collaboration, ACT.* The Atacama Cosmology Telescope: A Catalog of $>$ 4000 Sunyaev-Zel'dovich Galaxy Clusters / ACT collaboration. — 2020.

8. *E. S. Rykoff E. Rozo, M. T. Busha C. E. Cunha A. Finoguenov A. Evrard J. Hao B. P. Koester A. Leauthaud B. Nord M. Pierre R. Reddick T. Sadibekova E. S. Sheldon R. H. Wechsler.* redMaPPer I: Algorithm and SDSS DR8 Catalog / M. T. Busha C. E. Cunha A. Finoguenov A. Evrard J. Hao B. P. Koester A. Leauthaud B. Nord M. Pierre R. Reddick T. Sadibekova E. S. Sheldon R. H. Wechsler E. S. Rykoff, E. Rozo. — 2013.

9. *A. Saro S. Bocquet, E. Rozo B. A. Benson J. Mohr E. S. Rykoff M. Soares-Santos L. Bleem S. Dodelson P. Melchior F. Sobreira V. Upadhyay J. Weller T. Abbott F. B. Abdalla S. Allam R. Armstrong M. Banerji.* Constraints on the Richness-Mass Relation and the Optical-SZE Positional Offset Distribution for SZE-Selected Clusters / E. Rozo B. A.

Benson J. Mohr E. S. Rykoff M. Soares-Santos L. Bleem S. Dodelson P. Melchior F. Sobreira V. Upadhyay J. Weller T. Abbott F. B. Abdalla S. Allam R. Armstrong M. Banerji A. Saro, S. Bocquet. — 2015.

10. *Th. Boller M.J. Freyberg, J. Truemper F. Haberl W. Voges K. Nandra*. Second ROSAT all-sky survey (2RXS) source catalogue / J. Truemper F. Haberl W. Voges K. Nandra Th. Boller, M.J. Freyberg. — 2016.

11. *Olaf Ronneberger Philipp Fischer, Thomas Brox*. U-Net: Convolutional Networks for Biomedical Image Segmentation / Thomas Brox Olaf Ronneberger, Philipp Fischer. — 2015.

12. *Hengshuang Zhao Jianping Shi, Xiaojuan Qi Xiaogang Wang Jiaya Jia*. Pyramid Scene Parsing Network / Xiaojuan Qi Xiaogang Wang Jiaya Jia Hengshuang Zhao, Jianping Shi. — 2017.

13. *Abhishek Chaurasia, Eugenio Culurciello*. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation / Eugenio Culurciello Abhishek Chaurasia. — 2017.

14. *Xide Xia, Brian Kulis*. W-Net: A Deep Model for Fully Unsupervised Image Segmentation / Brian Kulis Xide Xia. — 2017.