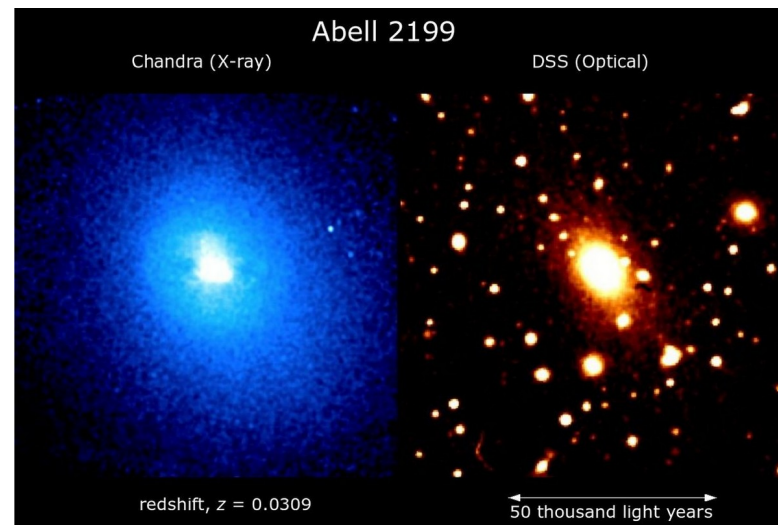


Нейросетевые методы поиска и сегментации объектов в данных современных космических обзоров

Научные руководители:
Герасимов С.В., к.ф.-м.н. Мещеряков А.В.
Студент: Немешаева Алиса,
4 курс бакалавриата ВМК МГУ

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 существует множество классических методов для поиска скоплений в различных данных.
- Скопления являются базовыми “кирпичиками”, из которых строится Вселенная. Изучая их параметры, можно исследовать структуру локальной Вселенной.

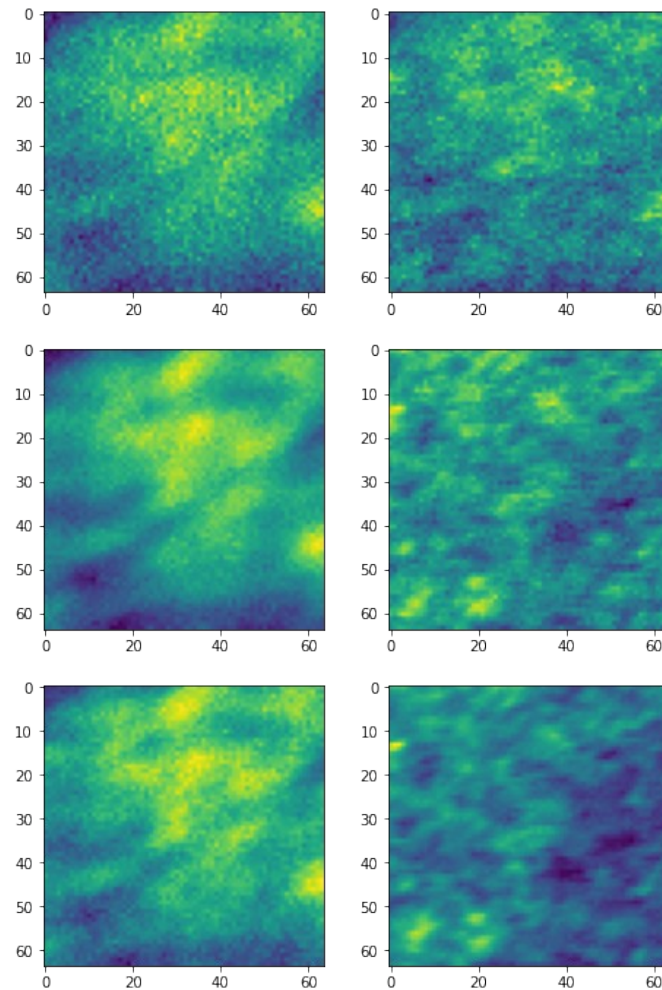


Введение

В данной работе будут использоваться данные микроволнового телескопа Planck в 6 каналах.

Данные этого телескопа покрывают всё небо.

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



Актуальность

Методы глубокого обучения дают следующие преимущества при анализе данных:

Стандартные алгоритмы сегментации усредняют информацию по нескольким каналам, в то время как с помощью нейросети можно охватить данные полностью.

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 данных нескольких каналов одновременно.

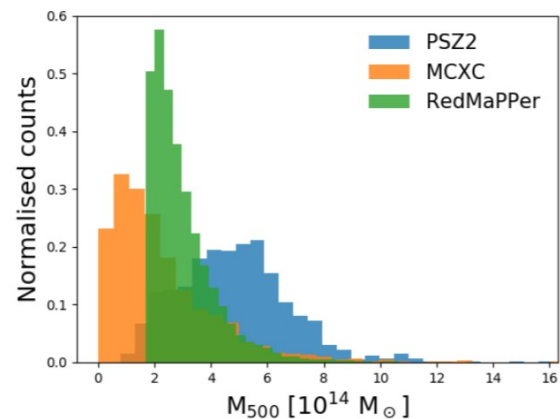
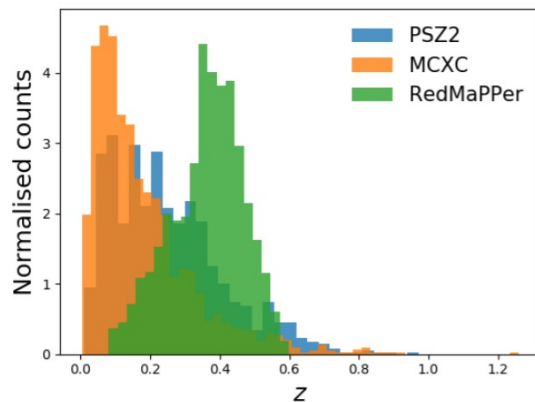
Постановка задачи

Сегментация, детекция и классификация скоплений галактик в многоволновых данных обзоров неба при помощи нейросетевых методов.

Обзор данных и обзор методов

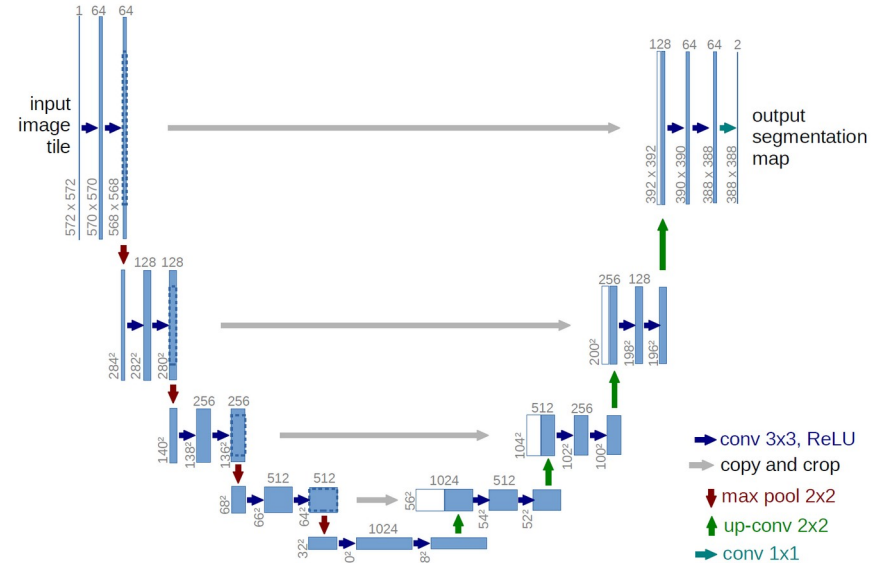
Существующие каталоги

- Рентгеновские:
 - eRosita
 - MCXC
- Микроволновые:
 - PSZ2
 - АСТ
- Оптические:
 - RedMaPPer



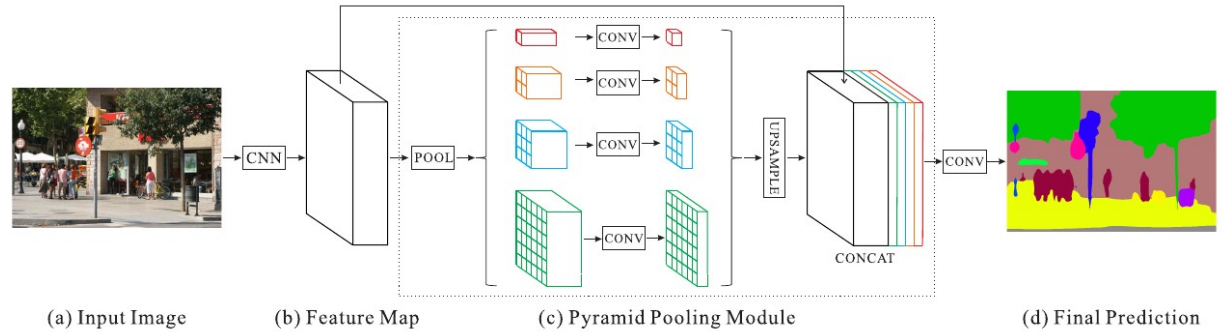
Обзор: U-net

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.

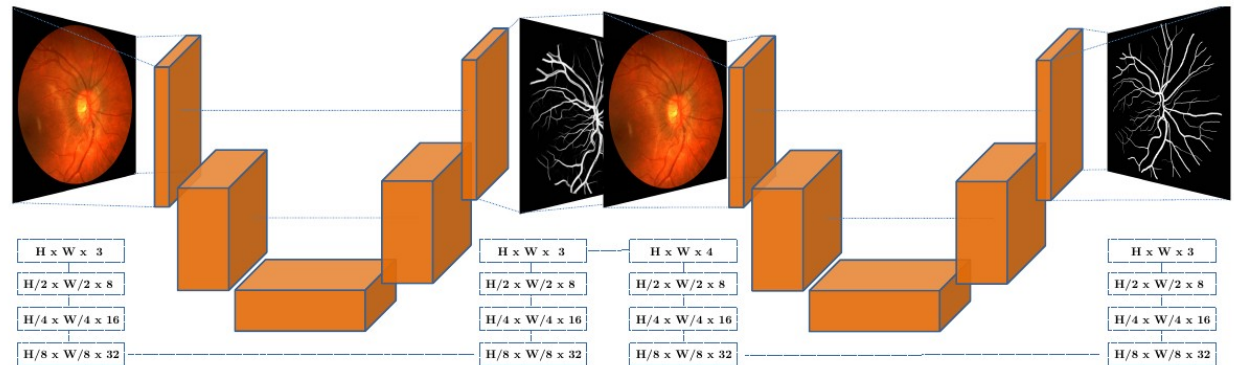


Другие нейросетевые модели

PSPNet



W-Net



Другие нейросетевые модели

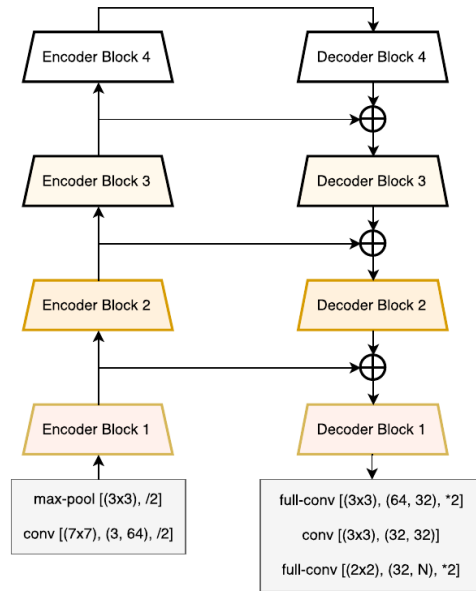


Fig. 1: LinkNet Architecture

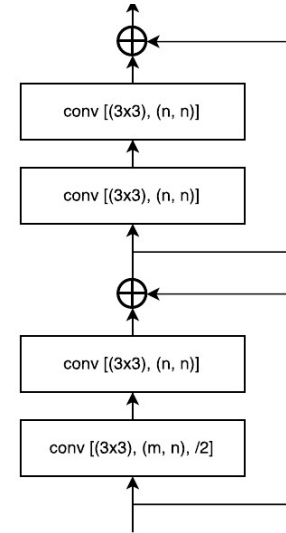


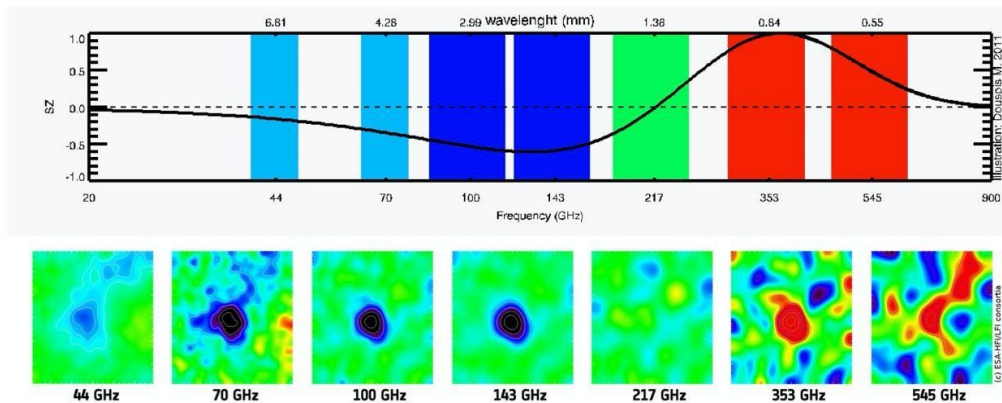
Fig. 2: Convolutional modules in *encoder-block (i)*

LinkNet

Данные Planck

Эффект Сюняева - Зельдовича — изменение интенсивности радиоизлучения реликтового фона из-за обратного эффекта Комптона на горячих электронах межзвёздного и межгалактического газа.

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.



Обзор: “Детекция эффекта Сюняева-Зельдовича”

<https://www.aanda.org/articles/aa/pdf/2020/02/aa36919-19.pdf>

Автор этой статьи использует для сегментации данных архитектуру U-net. Основной целью описываемой работы являлось создание алгоритма для детекции источников через эффект Сюняева-Зельдовича по данным телескопа Planck. Кроме самих обзоров неба, полученных телескопом, использовались еще три каталога скоплений для создания целевых данных:

- PSZ2 → `planck_z`, `planck_no_z`
- MCXC → `mcxcwpr`
- RedMaPPer → `rm30`, `rm50`

В открытом доступе не существует каталогов, полученных автором этой статьи, поэтому полезно повторить эксперимент, чтобы получить самостоятельно эти каталоги, а также сравнить результаты с новым микроволновым каталогом АСТ, которого на момент написания этой статьи ещё не было.

Построение решения

- 1) Предобработка.
- 2) Генерация данных для обучения.
- 3) Создание модели.
- 4) Обучение модели, подбор параметров модели.
- 5) Детекция и подбор параметров детекции.

Обработка данных

Обработка данных

- Данные Planck в разных каналах сильно отличаются по диапазонам значений. Чтобы улучшить результаты обучения, их нужно нормализовать, но так, чтобы можно было выделить значения, сильно отличающиеся от остальных.

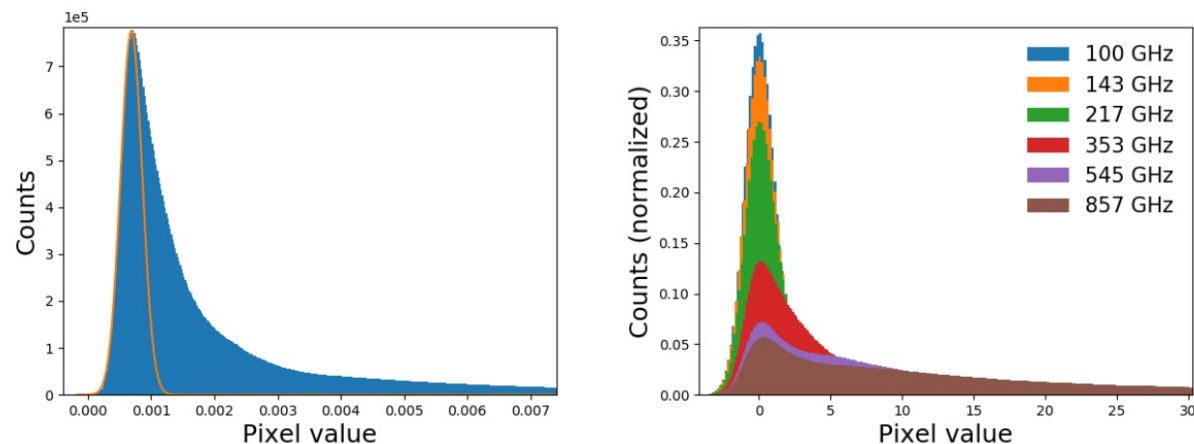
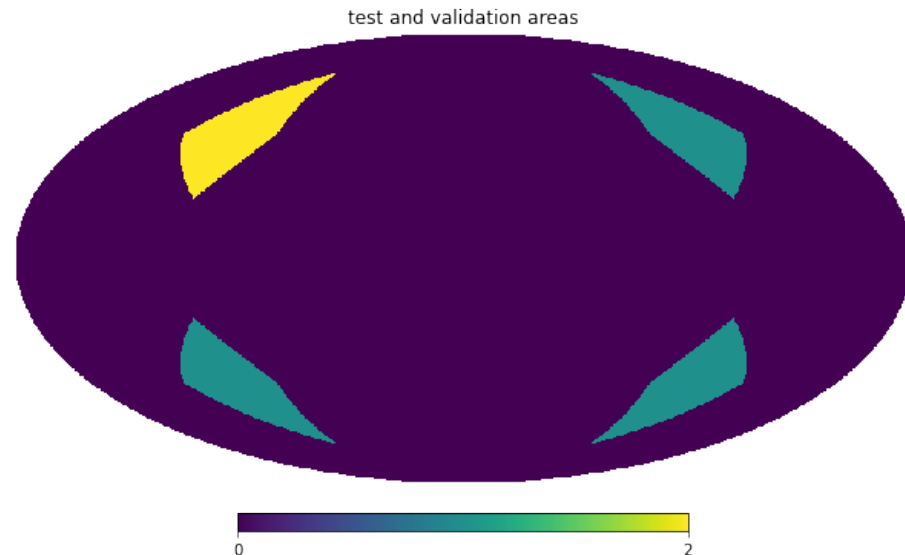


Fig. 2. Illustration of the data pre-processing. *Left:* pixel distribution of the map at 353 GHz. A Gaussian is fitted in orange up to the statistical mode of the distribution. The mean and standard deviation of the fitted Gaussian are used to normalise the data. *Right:* pixel distribution after normalisation of the six *Planck* HFI frequency maps.

Проекция данных

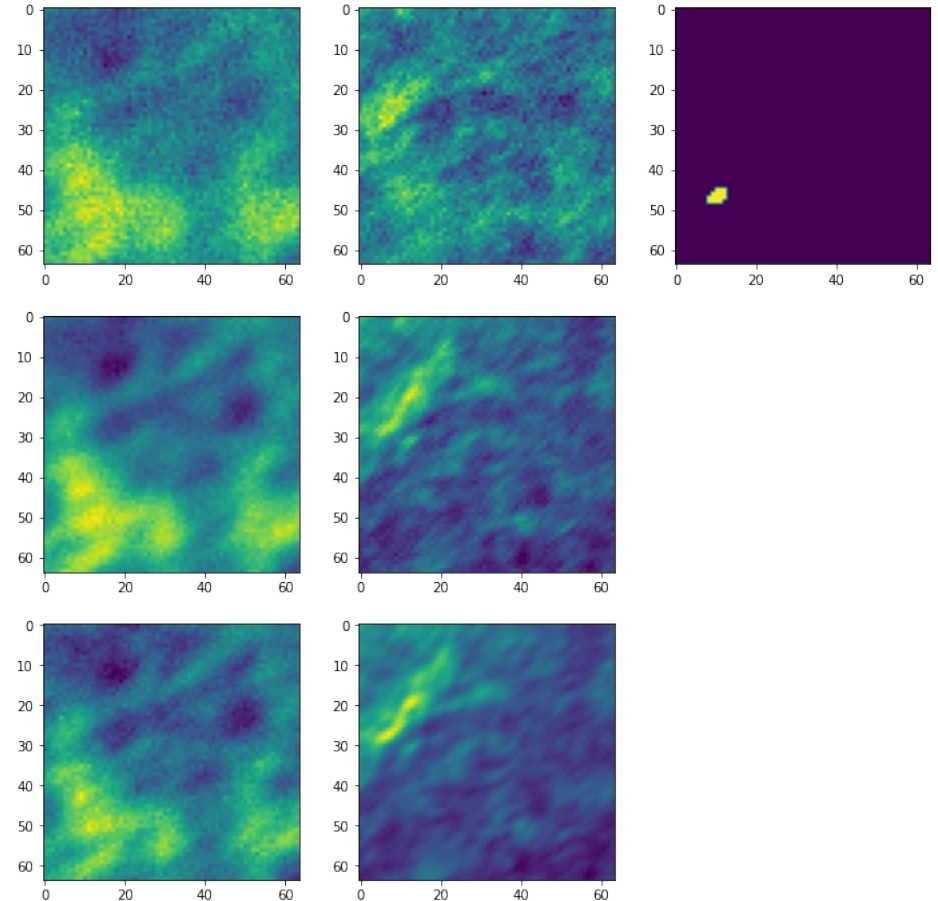
- Данные Planck хранятся в проекции HEALPix, в этой же проекции происходили обучение модели и детекция объектов.
- Разбиение с параметром $n_{\text{side}}=2$ позволяет получить 48 больших областей неба. Некоторые из них были использованы для тестирования полученной модели и для валидации, все остальные были использованы для обучения модели.



Генерация данных

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.

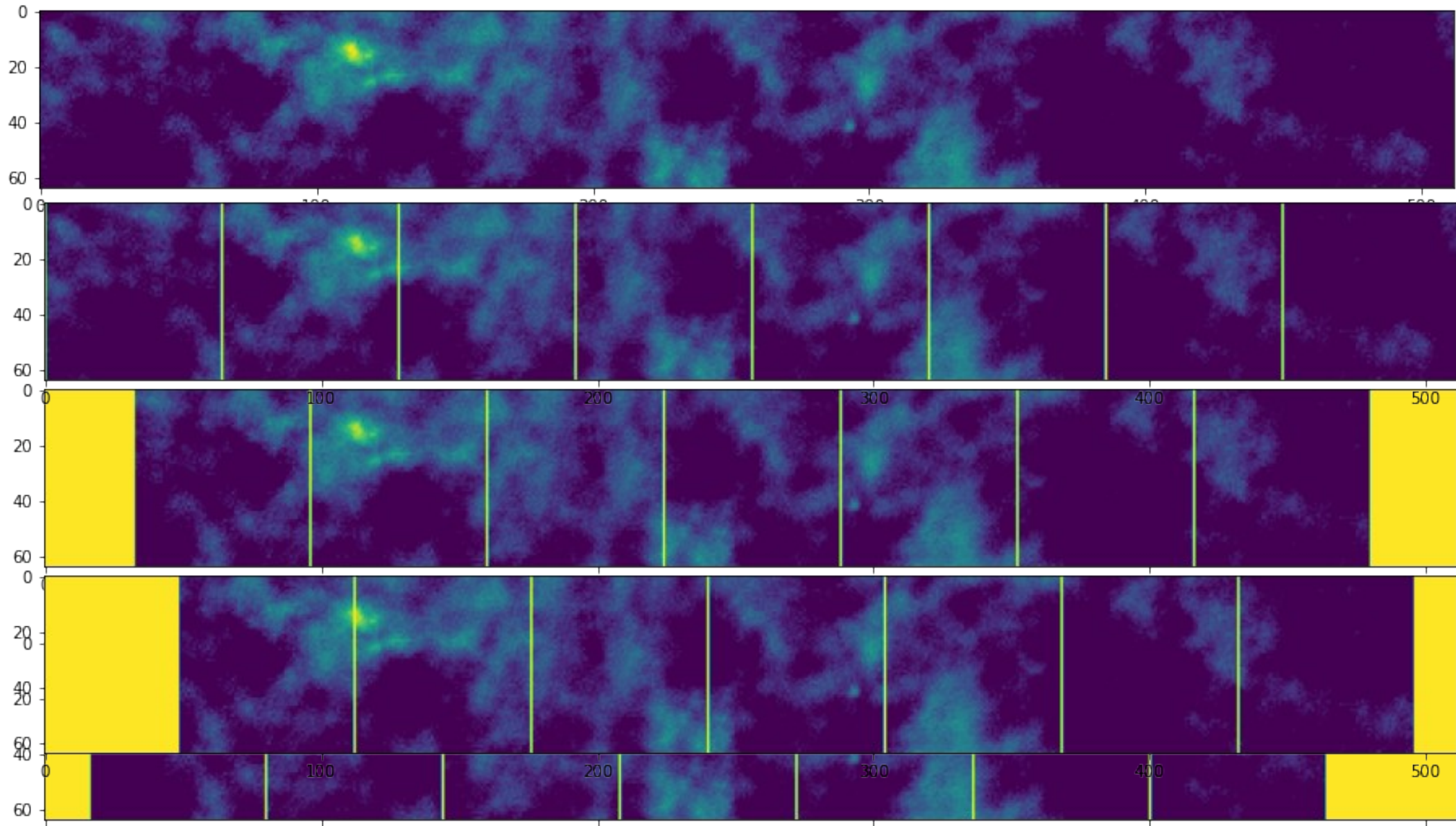
После этого 100000 патчей были использованы для обучения нейросети.



Детекция

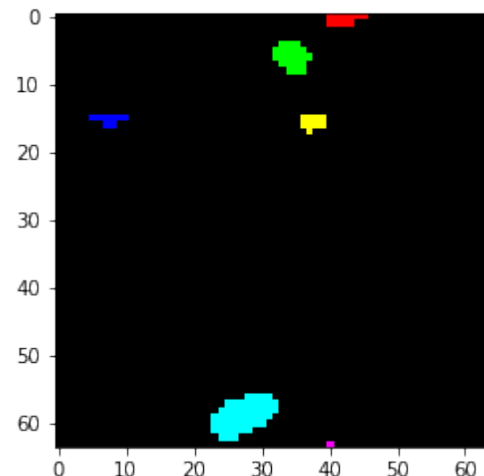
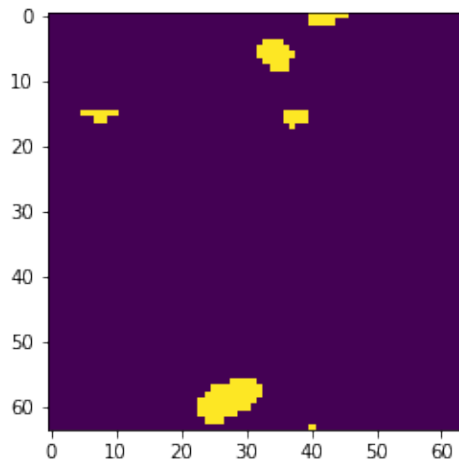
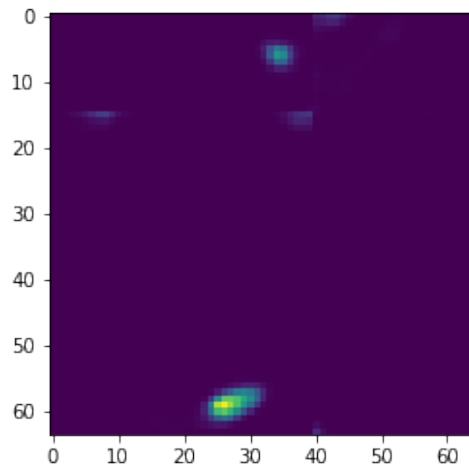
Детекция

- Чтобы полностью просканировать область неба, нужно разбить её на патчи размером 64x64, так же, как для обучения.
- Есть смысл сканировать некоторые данные повторно, так, чтобы разные части сканируемой области попадали в разные части патча.
- После сканирования полученные маски нужно объединить в общую prediction маску для всей сканируемой области неба.
- Таким образом появляется новый параметр детекции – “шаг”.



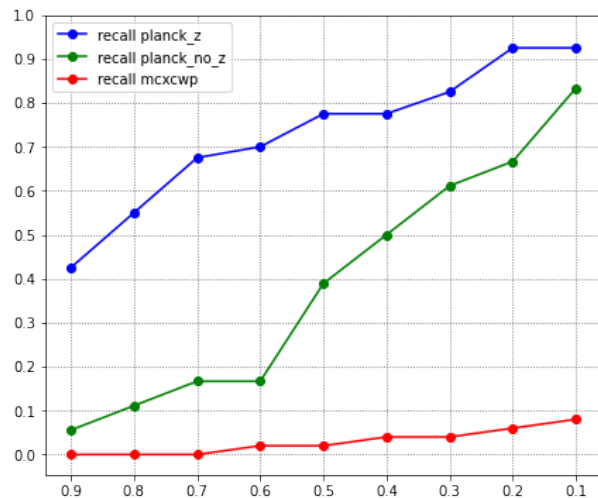
Детекция

- После получения маски сегментации для выбранной области неба выбирается порог детекции thr .
- На маске обнуляются пиксели, значение которых $< thr$.
- На маске “пятна” отделяются друг от друга.
- У каждого “пятна” находим барицентр.
- Координаты преобразовываем в Ra , Dec .

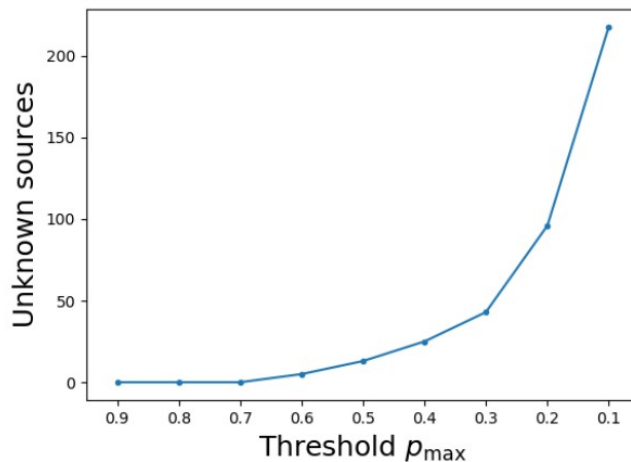
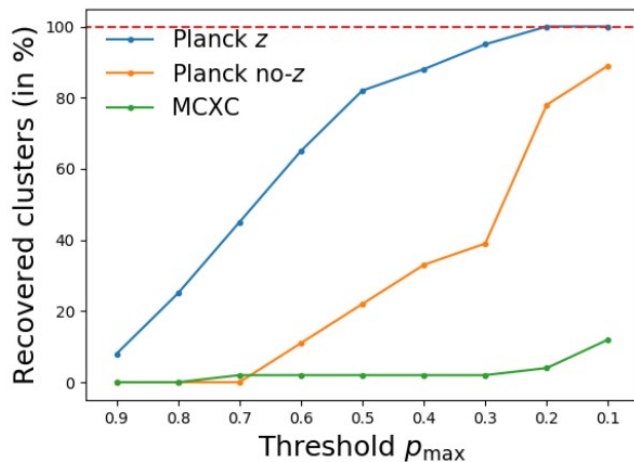
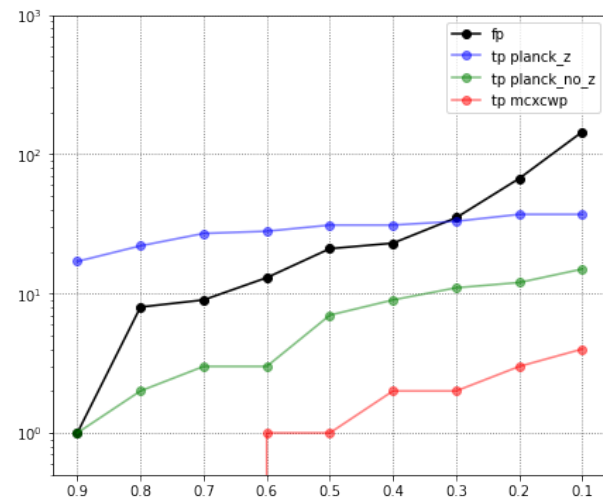
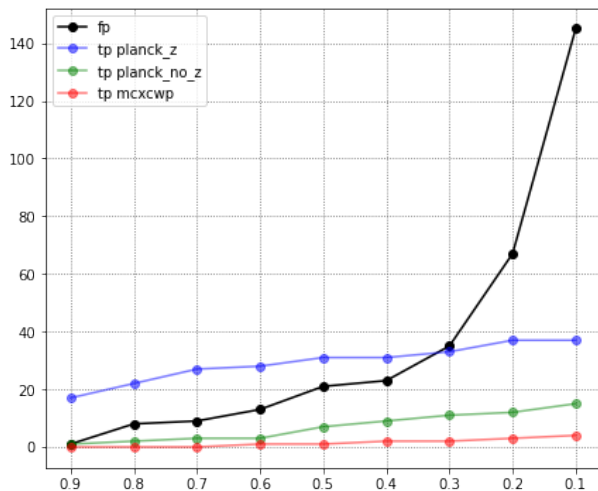


Сопоставление результатов с существующими каталогами

- Чтобы оценить, насколько хорошо получилось детектировать объекты, нужно сравнить свои результаты с существующими каталогами скоплений, рассматривая разные параметры детекции и обучения.

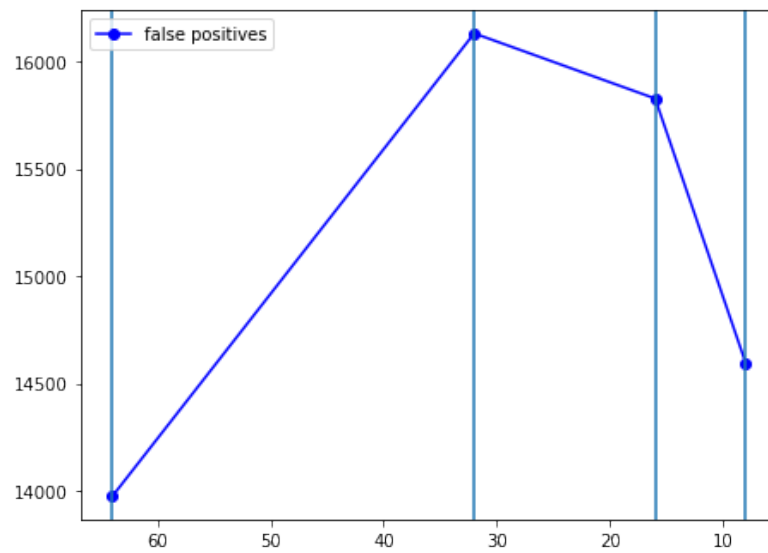
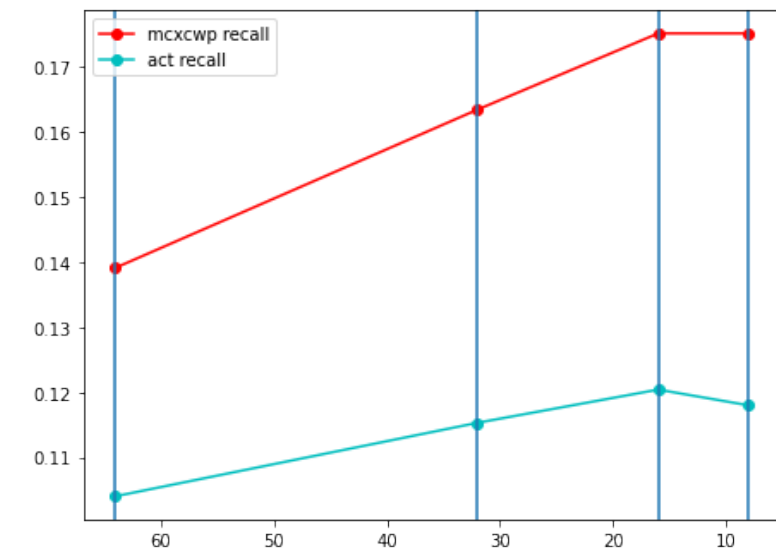
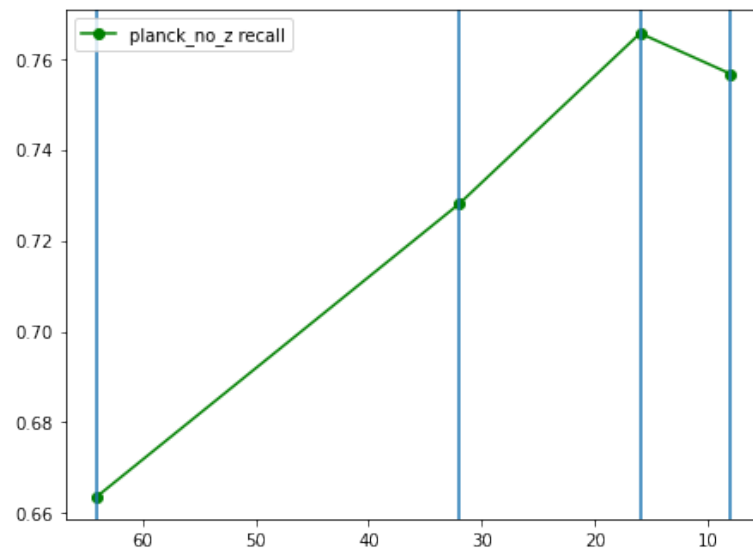
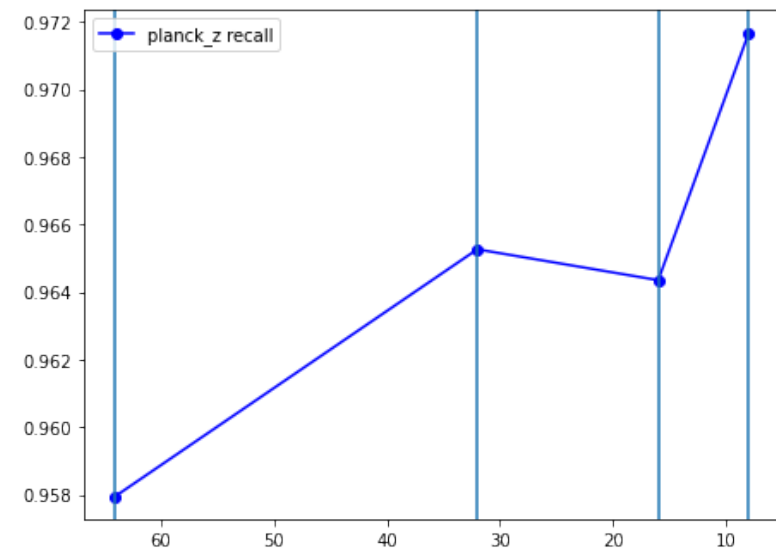


A&A 634, A81 (2020)



Количество неизвестных источников получилось на порядок ниже, отклик по planck_z ниже

Fig. 3. Results on the test area containing 40 Planck_z, 18 Planck_no-z, and 50 MCXCwP galaxy clusters. *Left:* Galaxy clusters recovered with different detection thresholds p_{\max} . *Right:* number of sources recovered with the U-net that do not belong to the Planck or the MCXC catalogue as a function of the threshold p_{\max} .



При уменьшении шага окна сканирования (step) мы наблюдаем увеличение полноты (recall) и уменьшение количества ложных объектов (fp) в ~1.132 раза.

Результаты

Результаты

- Создан каталог с параметрами:
 - 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

RA	DEC	area	min_rad	max_rad	mean_rad	min_pred	max_pred	tRA	tDEC	status	catalog
259.732221	5.094132	4.0	0.701176	2.119449	1.320277	0.901549	0.933698	NaN	NaN	fp	NaN
283.542185	26.460609	4.0	0.700159	2.110640	1.317352	0.901048	0.957279	283.519410	26.482998	tp	planck_no_z
280.308927	30.912179	7.0	0.700308	2.503611	1.673052	0.900285	0.988940	280.278406	30.932247	tp	planck_z
276.363142	30.419430	7.0	0.292191	2.616653	1.576871	0.906883	0.985809	276.336046	30.431996	tp	planck_z
266.093042	32.971296	5.0	0.447466	2.275328	1.422385	0.903614	0.968407	266.067814	32.998643	tp	planck_z

Сравнение с каталогом eRosita

- По данным eRosita также составляются каталоги, с которыми можно сравнивать полученные результаты.
- Эти каталоги созданы без использования нейросетевых методов.
- Ancat и psz2 созданы по одним и тем же данным, но в ancat найдено 640 других объектов в каталоге eRosita.

Matrix of UNIQUE eRosita matches

	ancat	brcat	psz2	redmp
ancat	1704	1064	574	676
brcat	1064	1715	531	641
psz2	574	533	593	227
redmp	680	643	228	4461

Текущие результаты

Созданы алгоритмы предобработки данных Planck.

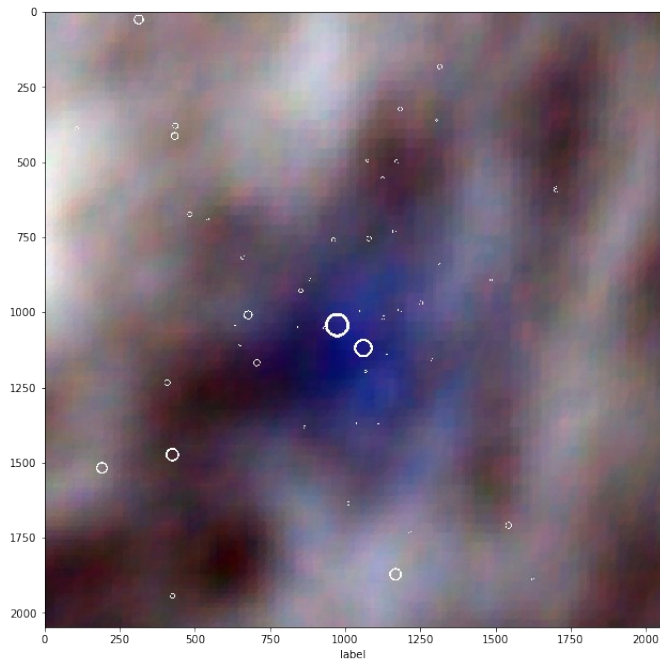
- Обучена модель для сегментации данных Planck.
- Созданы алгоритмы детекции масок сегментации, производимых моделью.
- Создан каталог с оптимальными (на данный момент) параметрами. 1704 объекта из этого каталога были найдены в каталоге eRosita.

Дальнейшие планы

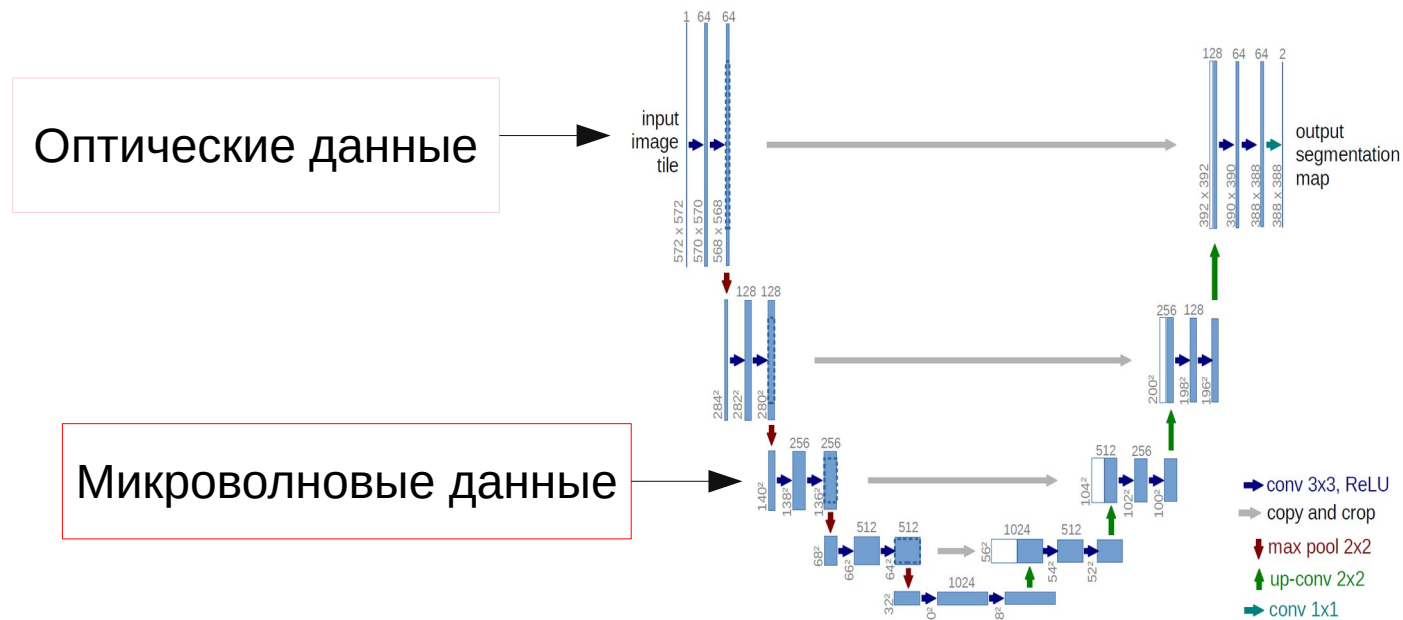
- Проверить, насколько влияют на детекцию меньшие значения шага.
- Обучить модель на каталоге АСТ.
- Перейти на детекцию в других диапазонах излучения.
- Классификация

Дальнейшие планы

- Готовы алгоритмы для предобработки оптических данных DESI LIS.
- Однако их разрешение выше, чем разрешение данных Planck, поэтому нужно придумать, как соединить эти каналы.



Дальнейшие планы



Один из вариантов: добавлять разные данные на разные блоки кодировщика U-net.