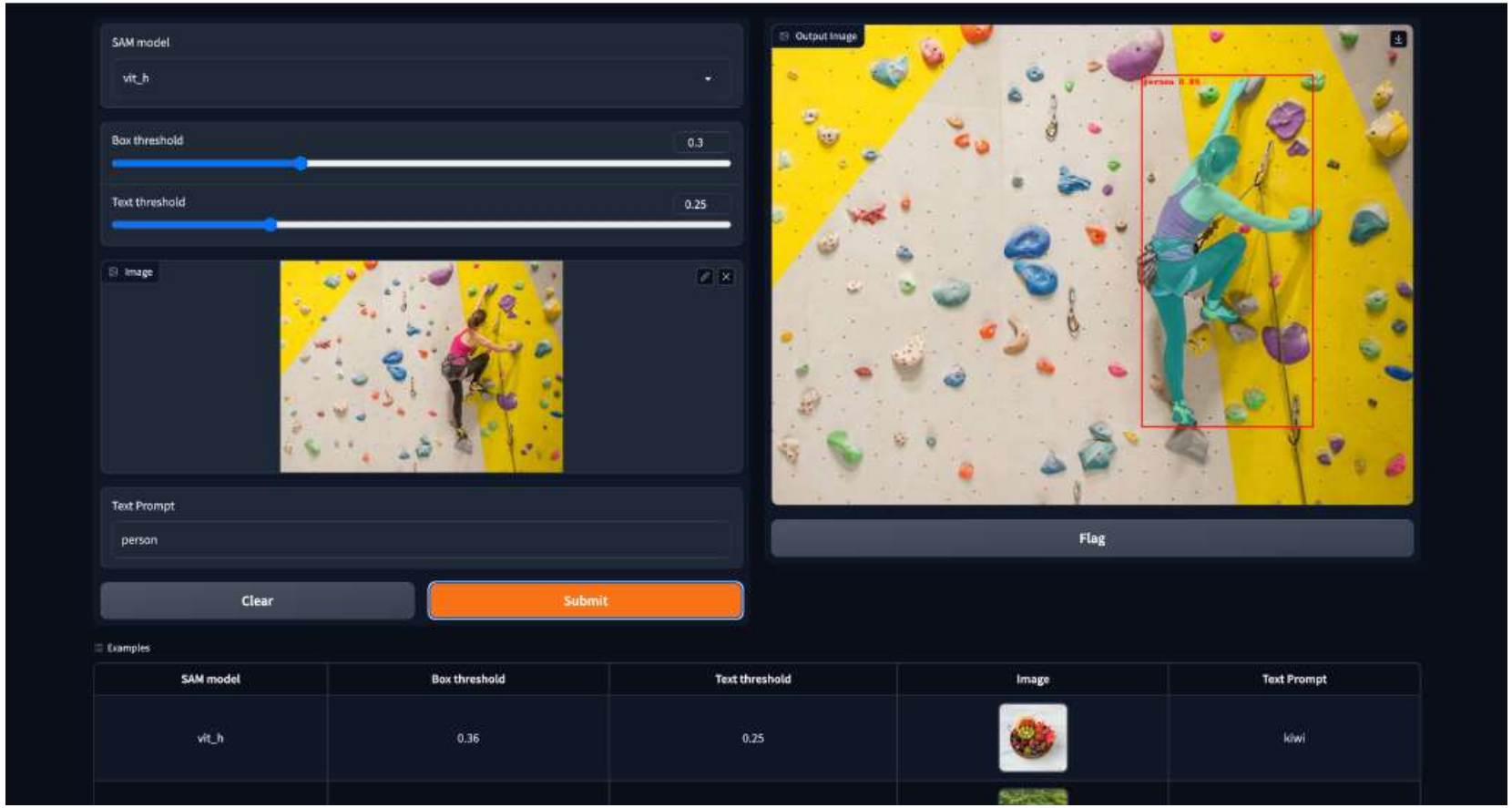
сегментация изображений по текстовому запросу для интеллектуального транспорта

СТУДЕНТ: КАЗАКОВА АНАСТАСИЯ, Б05-112

РУКОВОДИТЕЛЬ: ДМИТРИЙ АЛЕКСАНДРОВИЧ ЮДИН, ЦКМ

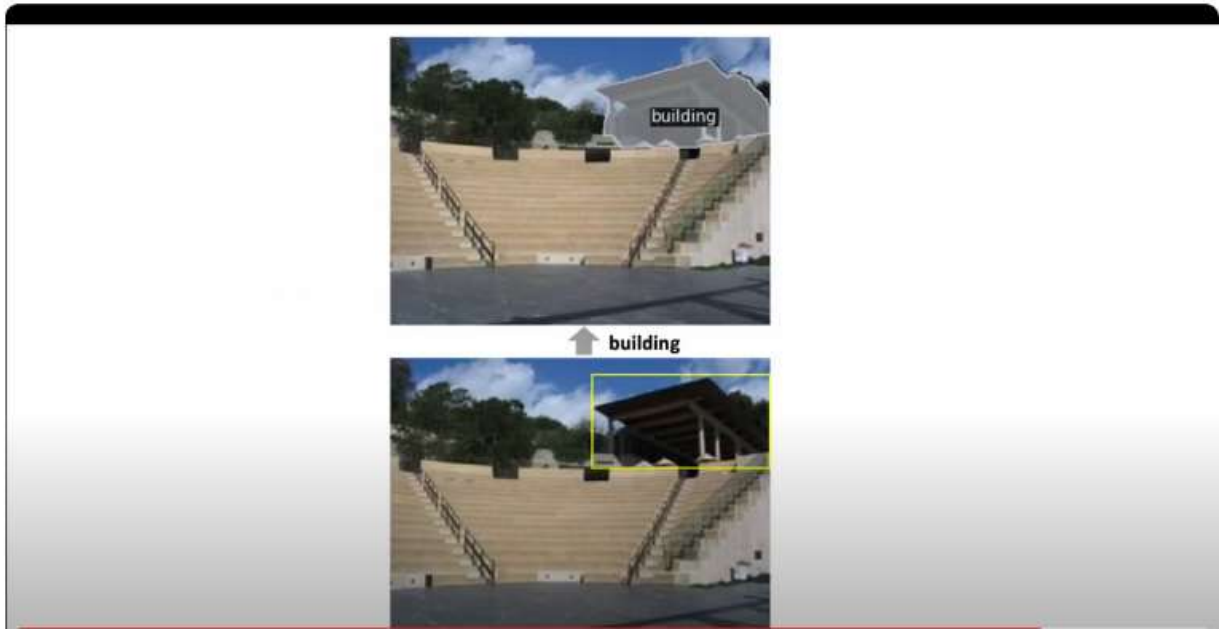
Задача сегментации объекта с текстовым промптом

- Сегментация изображений по классам - хорошо изученная задача.
- Решение задачи с открытым словарём, в которой мы сегментируем изображение при помощи текстового промпта – относительно новая задача.
- На вход подаётся промпт в виде текста и изображение
- На выход модель выдаёт маску
- Считается скор по метрике IoU



Мотивация

- Проект выполняется для расширения возможности систем помощи водителю беспилотных автомобилей и грузовой техники, интеллектуальных роботов.
- Нейросетевые методы распознавания данных могут обеспечить надежное распознавание целевых объектов. Вызовом является их надежное обучение и возможность работы в реальном времени.



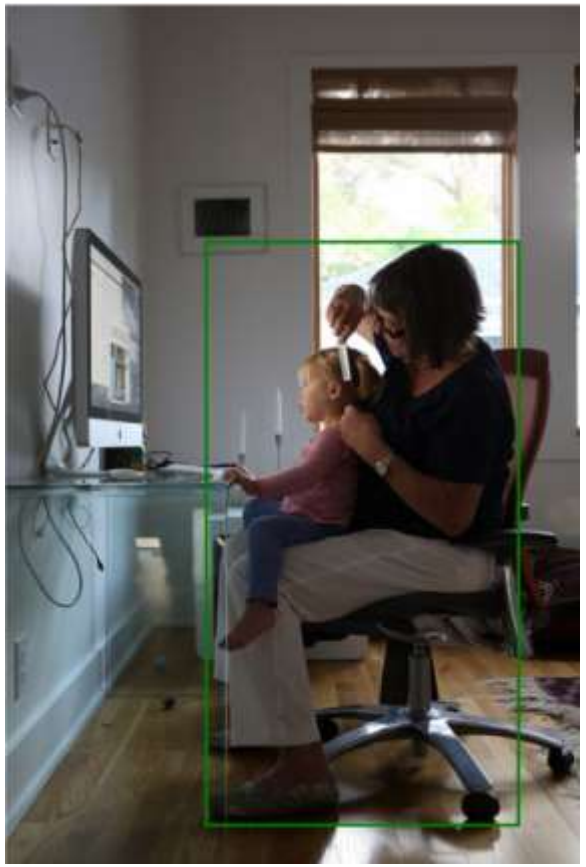
Метод

Наш метод заключается в создании двухстадийной модели. Первая часть - детектор, вторая -
модель сегментации.

- Задача детекции
  Для задачи детекции используются модели с возможностью детекции с подачей текстового промпта.
- Задача сегментации
  На вход подаём изображение, и bbox в качестве prompt, полученном на первом шаге, на котором можно наверняка сегментировать требуемый объект, так как он занимает почти всю площадь картинки bbox.

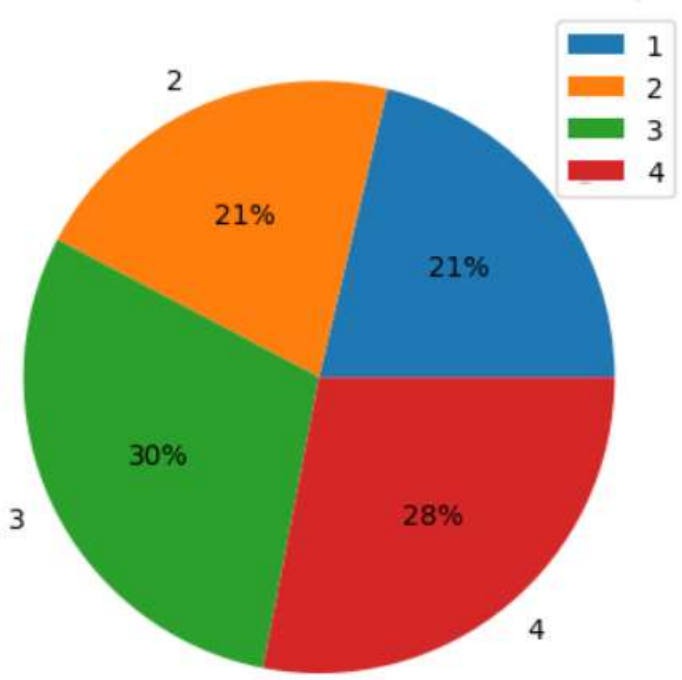
Inference YOLO-WORLD

Будем работать с датасетом RefCOCOg-val.
На YOLO-WORLD считаем inference на метрике IoU.
При валидации можно заметить, что модель не всегда правильно детектирует объекты на изображении. Разделим объекты на классы :
1. Модель ничего не задектировала
2. IoU == 0 - задектирован отдалённый объект
3. 0 < IoU < 0.5 - задектирован объект, который частично перекрывается содержится в таргетной bbox
4. 0.5 < IoU - условно правильно задектированный объект

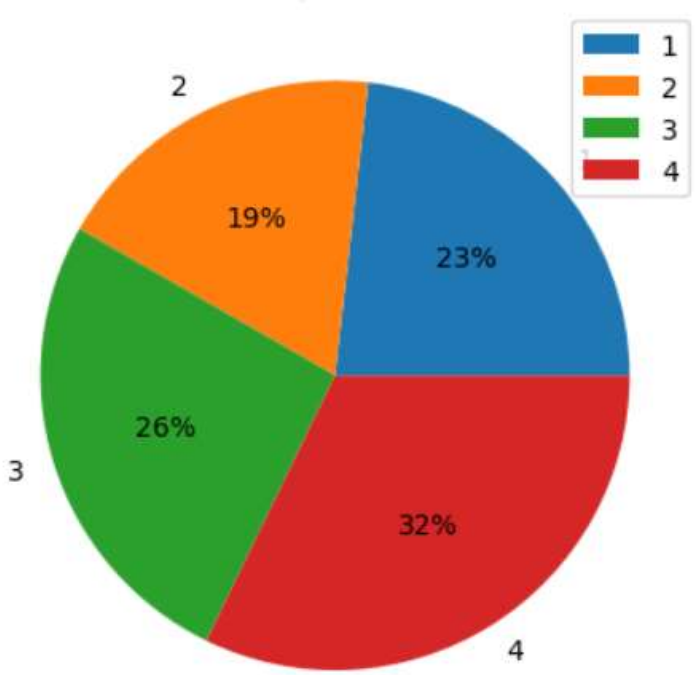


detected bbox для запроса : "child sitting on womans lap": IoU = 0.2634

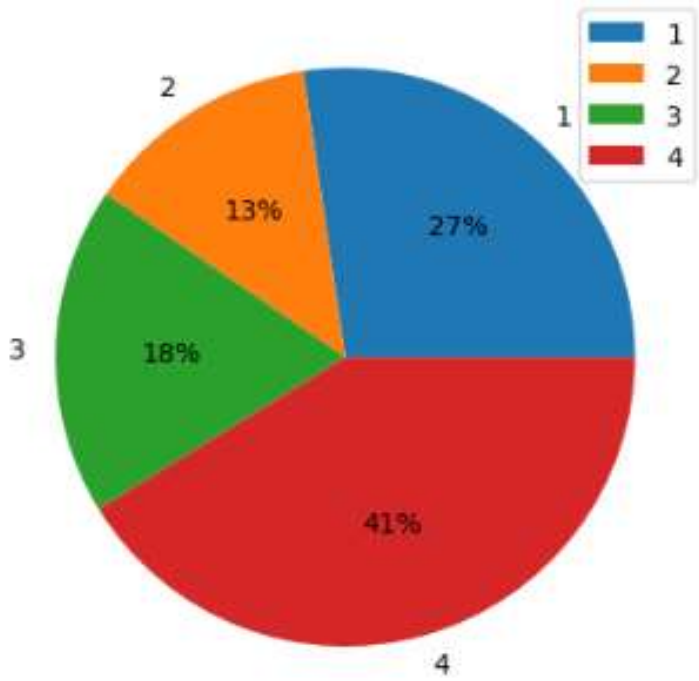
Inference YOLO-WORLD



Сложные запросы



Простые и сложные запросы



Простые запросы

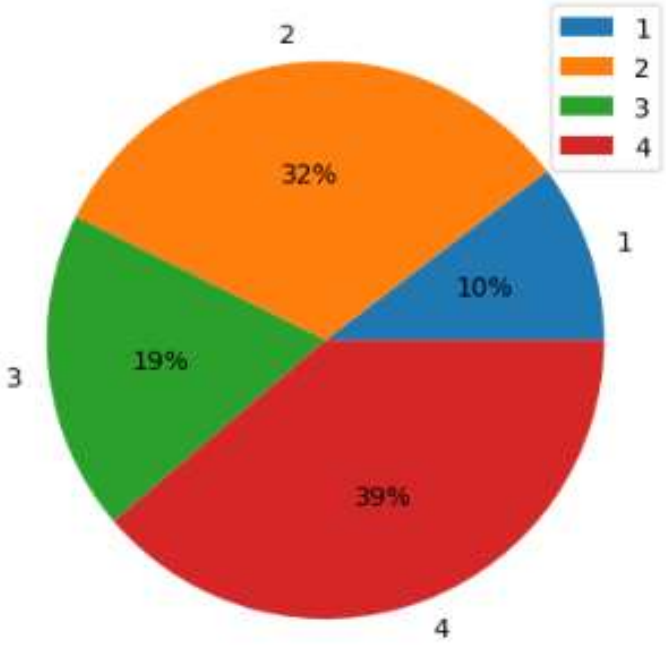
Inference YOLO-WORLD

Заметим с помощью диаграммы, что с использованием сложных запросов увеличивается число объектов класса 4 и уменьшается число объектов класса 3.
Из таблицы видно, что причина более низкой метрики на сложных запросах - менее точное определения положения задектированного объекта.
Это может быть связано с тем, что таргетный bbox получился включённый или, наоборот, включает в себя таргетный.

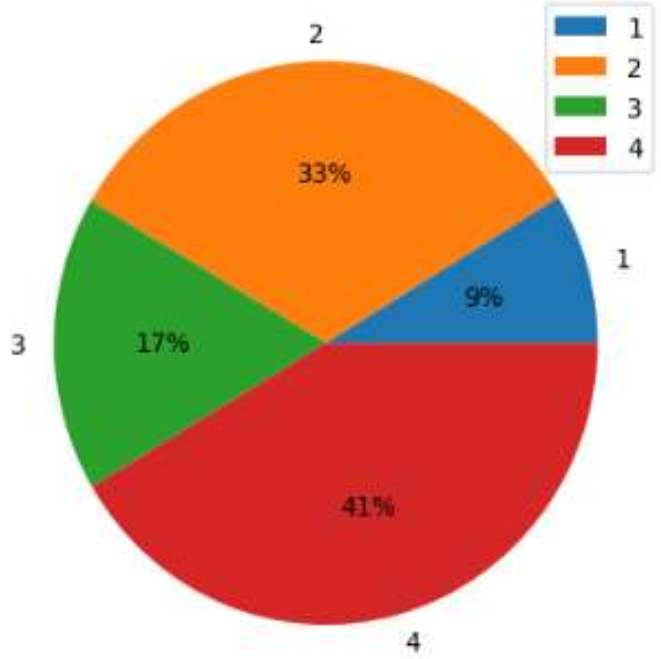
mAP	общий	2-4 классы	3-4 классы	3 класс	4 класс
все запросы	0.3136	0.4083	0.4363	0.3280	0.5240
сложные запросы	0.2995	0.3806	0.4056	0.3247	0.4911
простые запросы	0.3438	0.4731	0.5007	0.3393	0.5725

IoU	общий	2-4 классы	3-4 классы	3 класс	4 класс
все запросы	0.3277	0.4267	0.5623	0.1537	0.8932
сложные запросы	0.2950	0.3748	0.5114	0.1575	0.8856
простые запросы	0.3984	0.5482	0.6691	0.1480	0.9045

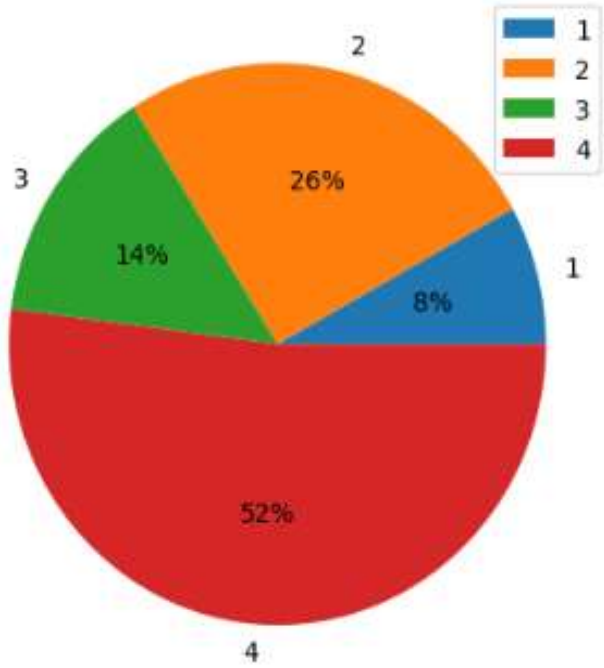
Inference Grounding Dino



Сложные запросы



Простые и сложные запросы



Простые запросы

Inference Grounding Dino

Метрики с использованием Grounding Dino выше за счёт:

- Меньшего числа незадектированных объектов
- Лучшего качества детекции задетектированных объектов

mAP	общий	2-4 классы	3-4 классы	3 класс	4 класс
все запросы	0.4099	0.4487	0.4937	0.3907	0.5357
сложные запросы	0.3739	0.4460	0.4611	0.3574	0.5108
простые запросы	0.4330	0.4578	0.5389	0.4009	0.5890

IoU	общий	2-4 классы	3-4 классы	3 класс	4 класс
все запросы	0.4460	0.4460	0.6974	0.1380	0.9257
сложные запросы	0.3832	0.4274	0.6687	0.1407	0.9216
простые запросы	0.5194	0.5882	0.7491	0.1781	0.9445

SAM results

IoU	общий	2-4 классы	3-4 классы	3 класс	4 класс
все запросы GD	0.2245	0.3143	0.5175	0.3389	0.7789

- Модель показывает хорошие результаты на объектах, которые были задетектированы на первом этапе

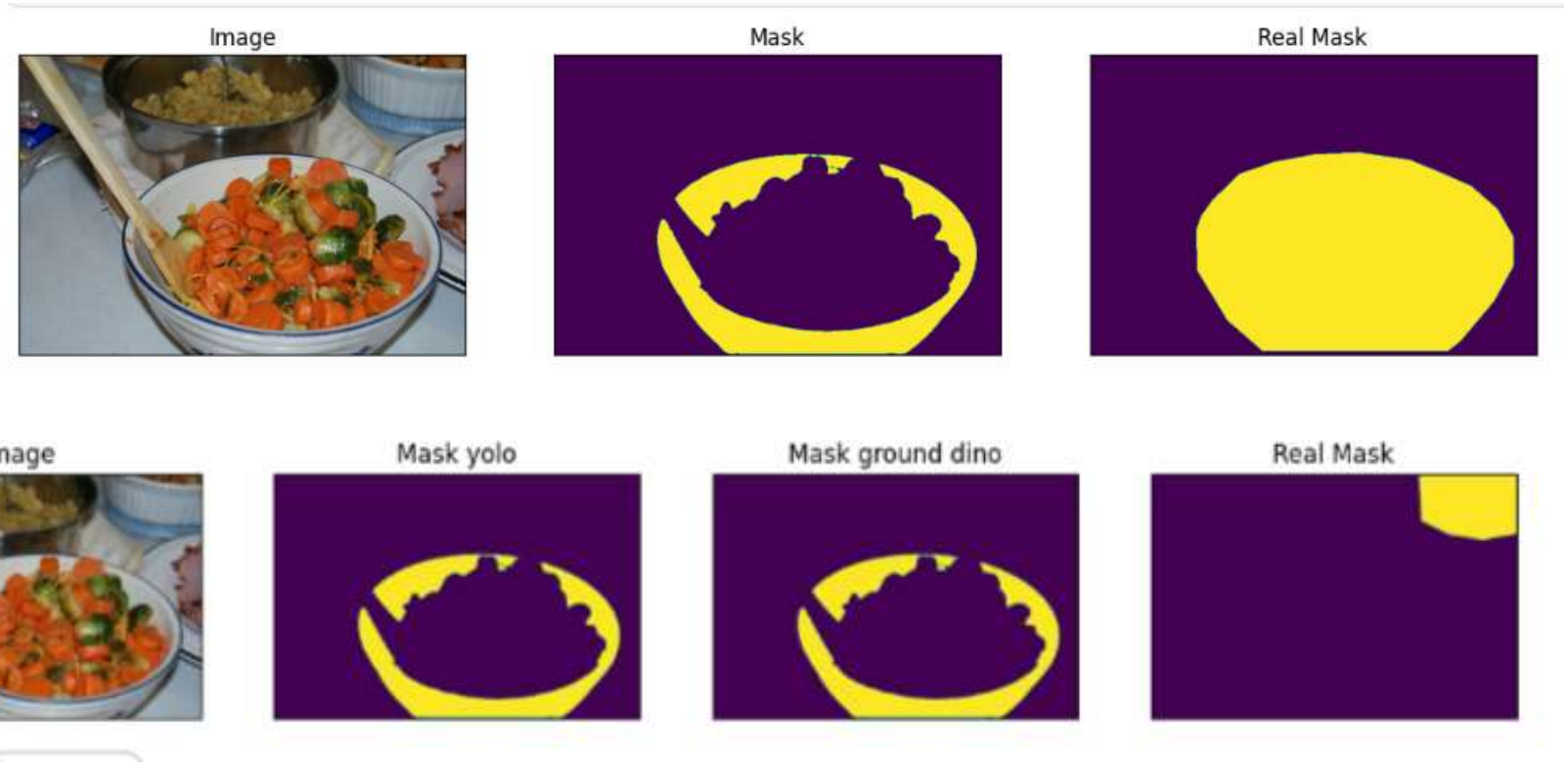
Задачи при выборе модели первого этапа:

- Уменьшение числа объектов первого класса
- Перевод объектов из 3 в 4 класс

1	UniLSeg-100	79.27		✓	Universal Segmentation at Arbitrary Granularity with Language Instruction		2023
2	UniLSeg-20	78.41		✓	Universal Segmentation at Arbitrary Granularity with Language Instruction		2023
3	GROUNDHOG	74.1		✓	GROUNDHOG: Grounding Large Language Models to Holistic Segmentation		2024
4	GLEE-Pro	72.9		✓	General Object Foundation Model for Images and Videos at Scale		2023
5	PolyFormer-L	69.2	71.15	✓	PolyFormer: Referring Image Segmentation as Sequential Polygon Generation		2023
6	PolyFormer-B	67.76	69.36	✓	PolyFormer: Referring Image Segmentation as Sequential Polygon Generation		2023
7	MagNet	65.36		×	Mask Grounding for Referring Image Segmentation		2023
8	X-Decoder (Davit-d5)	64.6		✓	Generalized Decoding for Pixel, Image, and Language		2022
9	VLT (Swin-B)	63.49		×	VLT: Vision-Language Transformer and Query Generation for Referring Segmentation		2022
10	LAVT	61.24		×	LAVT: Language-Aware Vision Transformer for Referring Image Segmentation		2021
11	VLT (Darknet53)	52.99		×	Vision-Language Transformer and Query Generation for Referring Segmentation		2021

SAM results

- Иногда SAM ошибается, как например здесь.
- Во избежание этого ему следует передавать key points, которые будут являться точками с максимальной вероятности нахождени объекта в них.
- Это можно реалтзовать помощью SAM, передав данные точки как promp



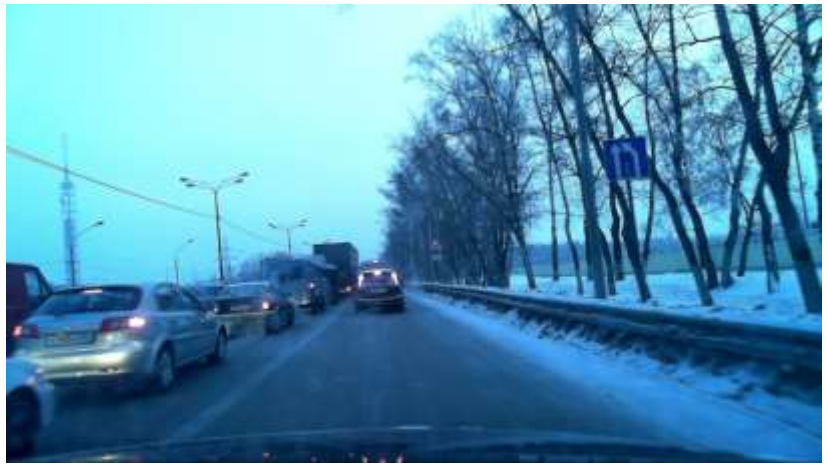
Интеллектуальный транспорт : мотивация

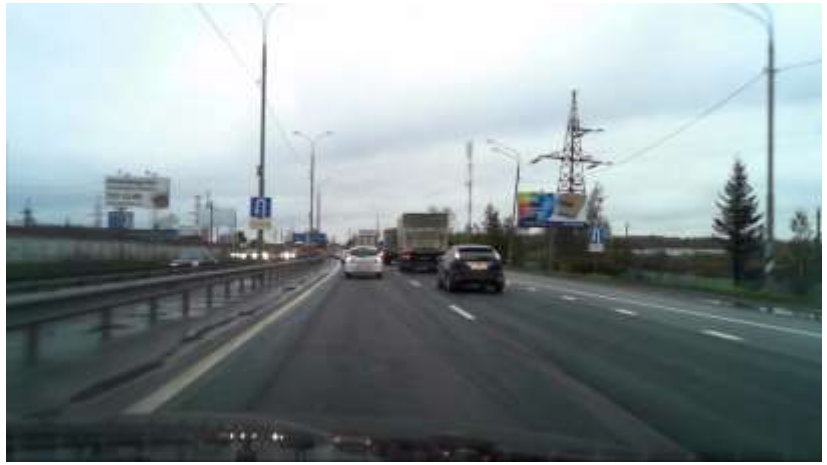
- refcocog - датасет, неспецифичный для задач интеллектуального транспорта.
- На нём можно было детально поработать с текстовыми промптами, и нам нужно перейти к нашей теме.
- Для этого нам потребуется датасет, специфичный для интеллектуального транспорта.
- Применим алгоритм к russian road dataset и увидим, как можно воспользоваться двухстадийностью модели для задач данного типа.

Подготовка датасета

Russian road dataset segmentation

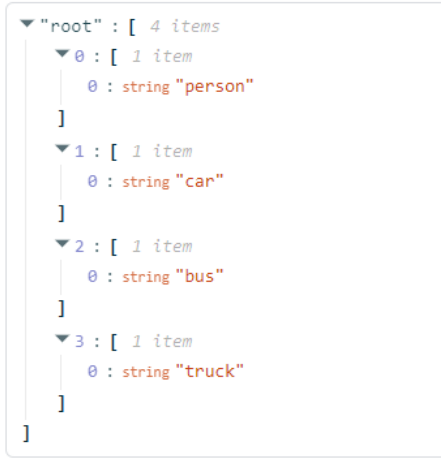
- Этот набор данных основан на В. И. Шахуро и А. С. Конушине, «Набор данных изображений российских дорожных знаков
- Авторы (Шахуро и Конушин) оригинального набора данных собрали около 100 000 изображений дорожных знаков, снятых с помощью автомобильного видеорегистратора. Фотографии были сделаны во время путешествия по России. Они уловили множество вариаций, искажений и дефектов. Этот набор данных был разработан для задачи обнаружения объектов, поэтому он содержит ограничивающие рамки и метки (типы дорожных знаков)
- Каждому файлу изображения *.jpg соответствует соответствующий файл *.json, содержащий маску дорожного знака.





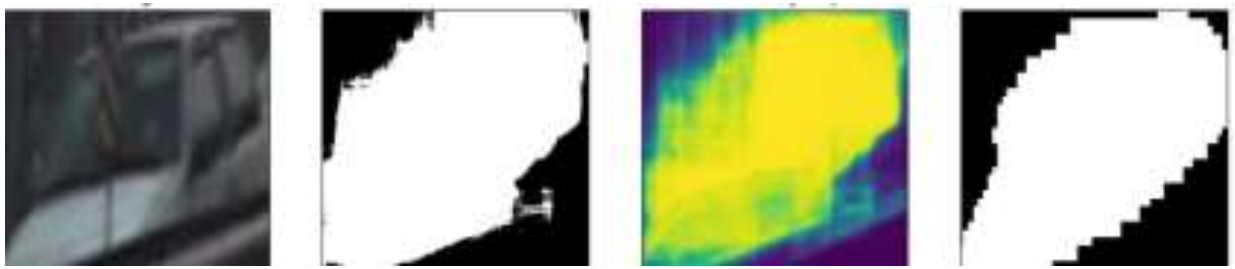
Создание кастомного датасета для модели детекции YOLO-WORLD

Преобразование датасета в coco format датасет с кастомными классами



SAM – segment anything model
Подготовка датасета

- Каждый элемент датасета – изображение объекта, который был задетектирован с помощью YOLO WORLD и его ground truth маска.
- На вход SAM подаю картинку, вырезанную по bbox. Получаю её маску на выход.



SAM + fine tuning

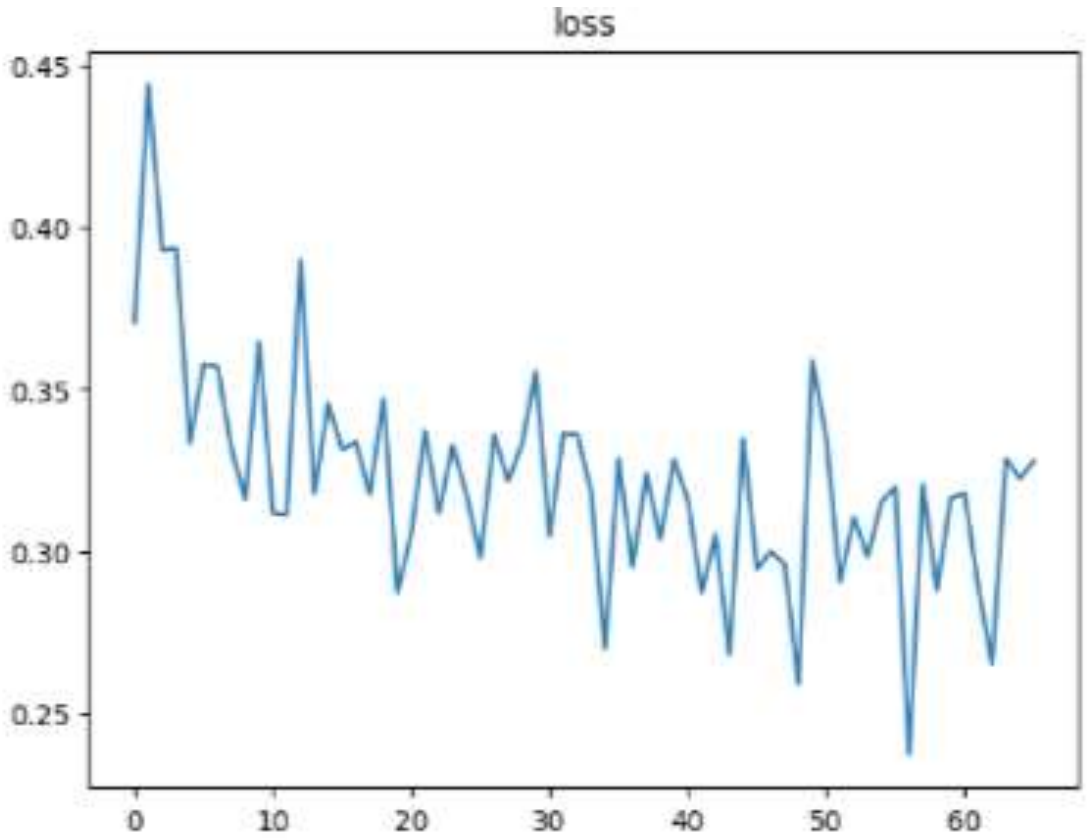
Iou без fine tuning 0.7035

После fine tuning 0.6789

Причины :

- В датасете большое число размытых изображений, на них лучше не обучать модель
- Небольшой набор train

Решение проблемы:
Iou после updated fine tuning 0.7342



SAM + fine tuning + augmentations

Решение проблемы:
IoU после updated fine tuning 0.7342

```python
def process_dataset(dataset):
    included_images = []

    for image in dataset:
        bbox = image["bbox"]
        if should_include(bbox): # Площадь больше пороговой площади (например, 1000)
            prob = inclusion_probability(bbox)
            if random.uniform(0, 1) < prob:
                included_images.append(image)
                break  # Включаем картинку в конечный датасет только если ее площадь достаточно велика

    return included_images
```

- Вычисление площади
- Фильтрация рамок Удаляются рамки, площадь которых меньше 1000.
- Вероятность включения: Вероятность включения изображения в датасет пропорциональна площади оставшихся рамок.
- Обработка датасета: Итерация по всем изображениям в датасете. Для каждого изображения проверяются его рамки. Если хотя бы одна рамка проходит фильтрацию, изображение добавляется в итоговый датасет с вероятностью, пропорциональной площади этой рамки.

Результаты

1. Разработана модель двухстадийной сегментации изображения с использованием текстового промпта
2. Обнаружены основные точки роста данной модели – работа с моделью на 1 этапе:

- Повышение числа детекций модели с помощью подбора модели первого этапа или её text / bbox threshold
- Повышение точности детекции – переведение объекта из класса 3 в класс 4

3. Предложен метод улучшения работы модели - key points на 2 этапе
4. Предложена и протестирована функция для умного fine tuning в outdoor задаче на 2 этапе

Ссылки

- https://www.kaggle.com/datasets/viacheslavshalamov/russian-road-signs-segmentation-dataset/data
  датасет Russian road
- https://paperswithcode.com/sota/referring-expression-segmentation-on-refcocog датасет refcocog