Task 2 - Object segmentation

Team 4:

- María José Millán
- Agustina Ghelfi
- Laila Aborizka

DATASET DESCRIPTION - KITTI MOTS

For this week we use the same dataset from the last week but using the segmentation annotations

Train	Val
12 sequnces	9 sequnces
8,073 pedestrian masks	3,347 pedestrian masks
18,831 car masks	8,068 car masks

Test
29 sequnces
No masks

train ← Images
val ←

train ← labels
val ←

COCO FORMAT

The annotations from train and validation is one .txt per sequence, structure:

`frame,id,class_id,width,height,rle`

id: have information of class_id performing floor division by 1000 and of intance_id by modulo 1000.

rle: compression method that encodes consecutive repeating values. Here is saved the bbox x,y,w,h.

MASK R-CNN OR MASK2FORMER

Both models need annotations with the coordinates of the bounding box and a dictonary of size and contours that contain the connected components

YOLO

Annotations normalized

```
class_id,x1,y1,x2,x2,.....xn,yn
```

Each line in the .txt file corresponds to a polygon. If an object has two connected components, there will be two separate lines, like in the example.

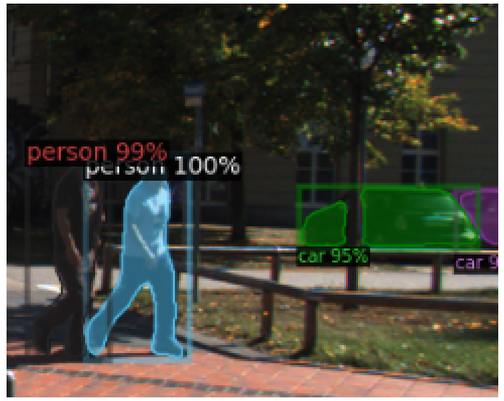
Annotations denormalized

```
{"images":{..}
 "annotations":{..,
        "bbox":[xmin,ymin,w,h],
        "segmentation":{ "size":[height,width]
                         "counts": rle},
        ....,..  }
 "categories":{..}}
```

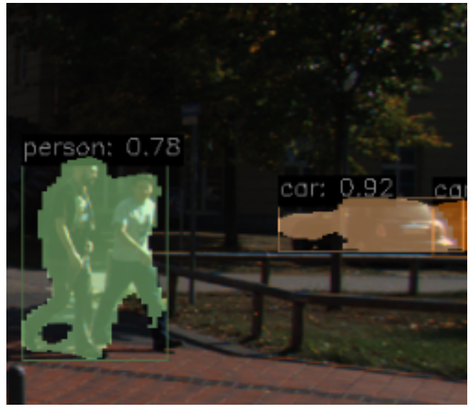
[1] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, A. Geiger, and B. Leibe, "MOTS: Multi-Object Tracking and Segmentation," arXiv preprint arXiv:1902.03604, 2019.

INFERENCES

MASK RCNN ✔

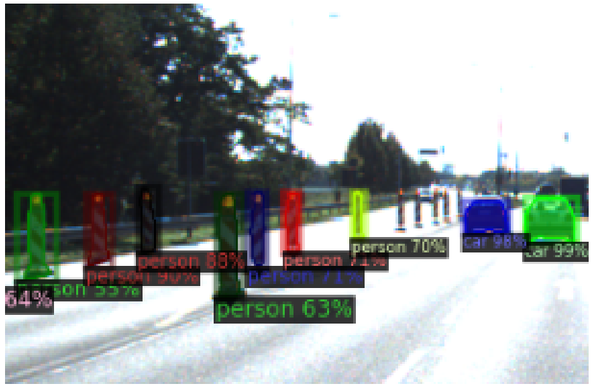


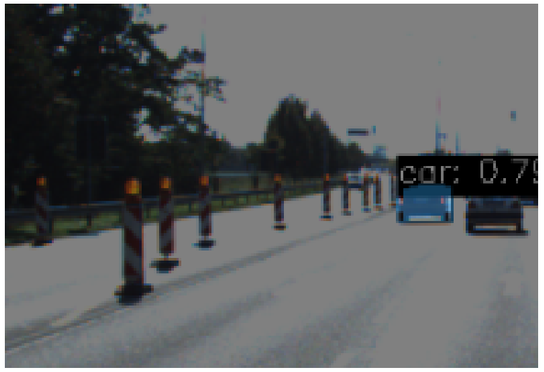
MASK2FORMER



YOLOV11X-SEG

The comparison of models was performed using the same **confidence threshold of 0.5.** The differences are clearly visible, in this specific example, Mask R-CNN has detected both people and cars with a highly precise segmentation mask. In the case of the car, we can clearly see two connected components since the model identifies the car as a whole but does not include the tree as part of it. On the other hand, in the other two models, the tree becomes part of the mask, and in Mask2Former, both people are detected as a single entity.





In this case, the focus is on highlighting false positives in person detection. With Mask R-CNN, all traffic cones are mistakenly detected as people. Moving on to the YOLO models, we can see that only some cones are detected as people (specifically, those closer to the camera). In Mask2Former, no people are detected at all, but it is worth noting that it only identifies a single car.

EVALUATION

Metric Used: mAP50-95

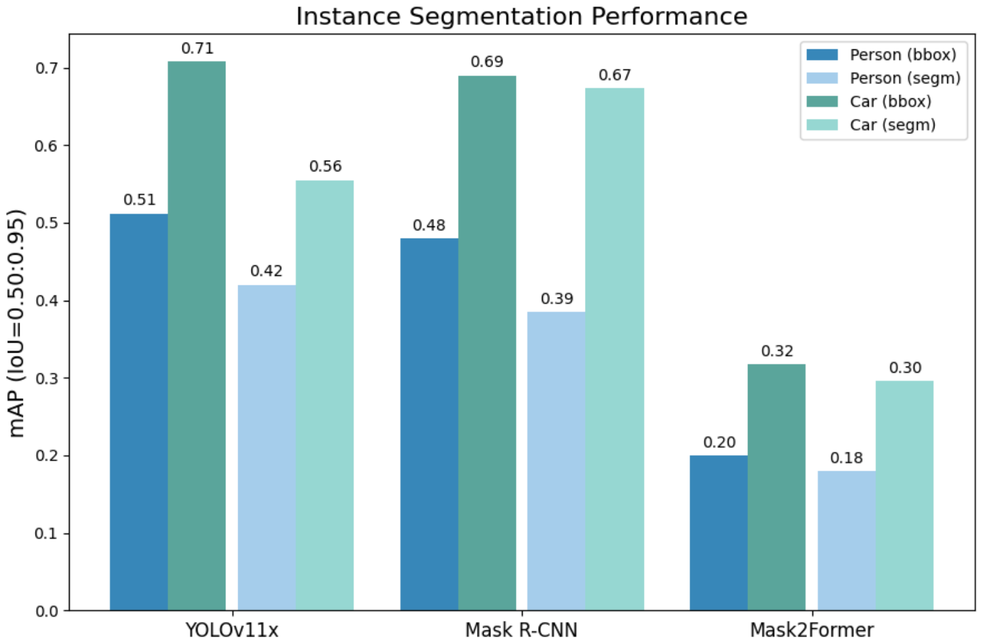
To evaluate the different models we used the Mean Average Precision (mAP) metric, with varying thresholds from 0.50 to 0.95 (mAP50-95).
On this week we analize this metric for bounding box detection and segmentation, because they provide complementary insights when comparing models.
The bounding box metric helps evaluate how well the model detects and localizes objects in the image while the segmentation metric assesses the model's ability to precisely outline the shape of the detected objects. Evaluating both aspects ensures a more comprehensive comparison of model performance.

Label Conversion

Since KITTI-MOTS uses different labels than COCO, a class mapping was performed. Like last week, depending on the model is the implementation of the conversion.

Quantitative Results

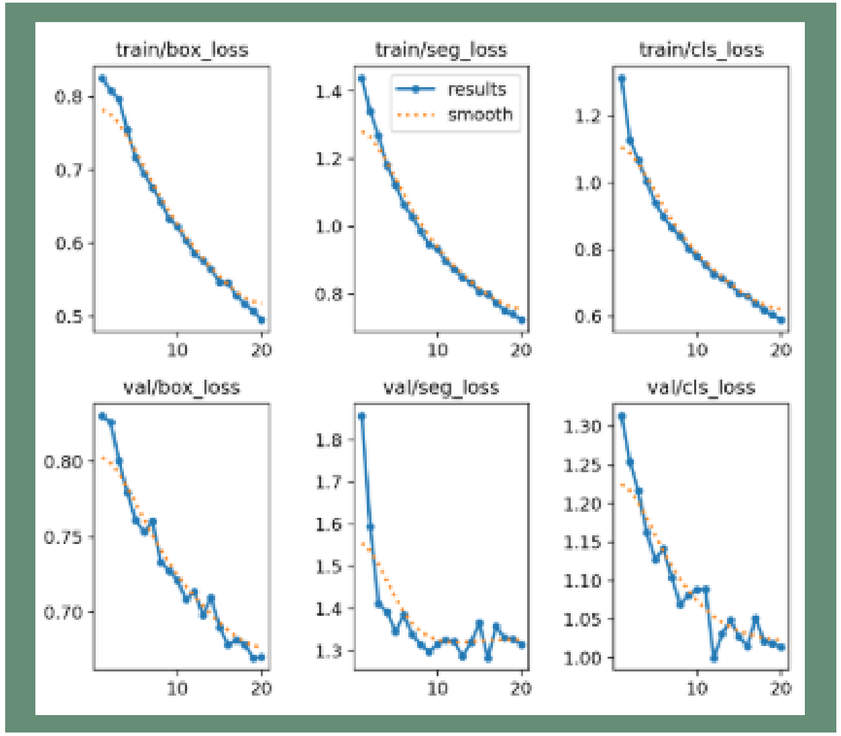
The obtained mAP50-95 values are shown in the barplot below, for each class.



Observations

- YOLOv11x-seg performs well in detecting persons and cars. Its segmentation performance is slightly higher than Mask R-CNN in the person class(0.42).
- Mask R-CNN outperforms YOLOv11k in car segmentation tasks, achieving the highest score in Car (0.67).
- Mask2Former lags behind in both bounding box and segmentation tasks, indicating that it may not be as effective in this specific instance segmentation benchmark.

FINETUNE-YOLOV11X-SEG

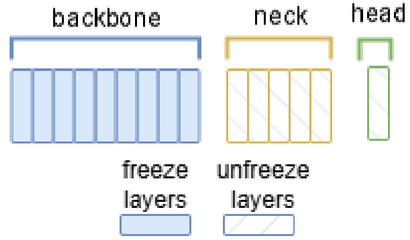


Graphs of the best model

The model is learning well, as indicated by the decreasing training and validation losses:

- box_loss: decrease steadily, showing that the model is improving in detecting object locations.
- seg_loss: decrease but with more fluctuations in validation loss, which might indicate some instability in segmentation performance.
- cls_loss: decrease significantly, which means the model is improving in correctly classifying objects

Best freeze configuration



Different hyperparameters such as learning rate, optimizer and momentum were analyze. Different parts of the model were frozen, and augmentations were applied in multiple combinations, we are presented some of them in the table.

Augmentations with best freeze configuration

	Pedestrian	Car
backbone and neck freeze lr=0.001	mAP50=0.66 mAP50-95=0.42	mAP50=0.75 mAP50-95=0.50
backbone freeze lr=0.001 with clf=1 and dlf=2	mAP50=0.7 mAP50-95=0.46	mAP50=0.76 mAP50-95=0.52
backbone freeze lr=0.0005 with clf=1 and dlf=2	mAP50=0.7 mAP50-95=0.46	mAP50=0.76 mAP50-95=0.52
hsv_h=0.01,hsv_s=0.2, hsv_v=0.4, translate=0.1, scale=0.4	mAP50=0.60 mAP50-95=0.36	mAP50=0.71 mAP50-95=0.44
hsv_s=0.3,translate=0.2 scale=0.4, crop_fraction=0.3	mAP50=0.68 mAP50-95=0.46	mAP50=0.77 mAP50-95=0.52
hsv_h=0.01,hsv_s=0.3, hsv_v=0.4, translate=0.2, scale=0.4	mAP50=0.60 mAP50-95=0.37	mAP50=0.72 mAP50-95=0.45

EVALUATION COMPARISONS

mAP are all from segemntation metric

	PRETRAINED	FINETUNE
Person	mAP50=0.66 mAP50-95=0.42	mAP50=0.70 mAP50-95=0.43
Car	mAP50=0.80 mAP50-95=0.55	mAP50=0.80 mAP50-95=0.57

FINETUNE-MASK R-CNN

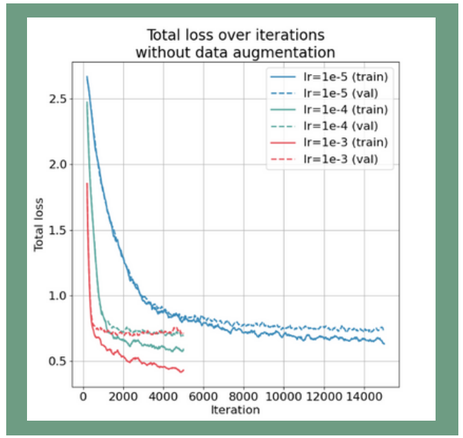
To improve the performance of Mask R-CNN on KITTI-MOTS, several experiments were conducted, including layer freezing, hyperparameter tuning, and data augmentation.

Hyperparameter Optimization

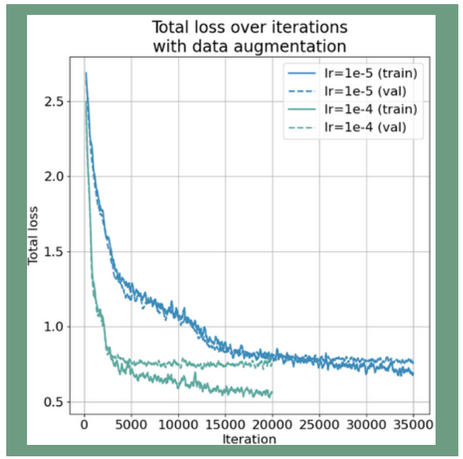
Experiments were conducted using AdamW and SGD optimizers, with learning rates of 1e-5, 1e-4 and 1e-3. SGD tests obtained better results. The graphs below present the variation of total loss in the case of SGD as the learning rate changes. Some experiments were stopped earlier due to poor convergence or overfitting.

The total loss in Mask R-CNN comprises:

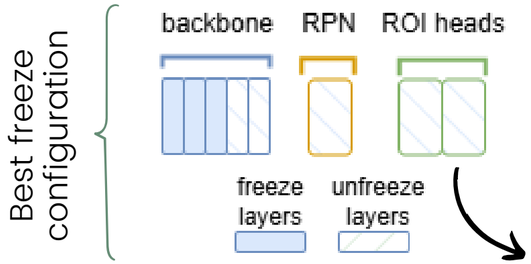
- Classification loss
  How well the model predicts object categories
- Bounding box regression loss
  Accuracy of bounding box coordinates.
- RPN loss
  Region proposal accuracy.
- Masks loss
  How well the model predicts object masks for detected objects.



Various augmentation strategies were tested using Albumentations, but no significant performance improvements were observed. The loss curves with and without augmentation show similar trends, suggesting that augmentation did not have a strong effect in this case.



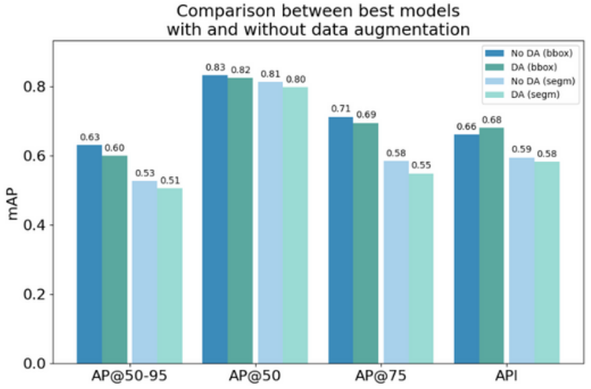
Layer Freezing Strategy



- The freezing process involved gradually unfreezing layers from the head to the entire network.
- Best configuration: freezing first three layers of the backbone, while keeping the rest trainable.

In Mask R-CNN, the ROI Heads also include the Mask Head, which is responsible for predicting segmentation masks for each detected object.

BEST EXPERIMENTS AND RESULTS



For the augmented case, there is a slight drop in accuracy, indicating that the current augmentation strategy may not be optimal.

Overall, the results suggest that further tuning or different augmentation techniques may be needed to achieve meaningful improvements.

		mAP@50-95 Person	mAP@50-95 Car
FREEZE_AT=3 FPN=🔒 RPN=🔒 ROI=🔒 SGD lr = 1e-5	No data augmentation	bbox: 0.51 segm: 0.38	bbox: 0.69 segm: 0.68
	Compose([A.OneOf([A.RandomCrop(width=1000, height=300), A.RandomCrop(width=500, height=150),]), A.Illumination(p=0.5, intensity_range=(0.01, 0.2))]) min_area = 200, min_visibility = 0.1	bbox: 0.51 segm: 0.35	bbox: 0.68 segm: 0.66

FINETUNE-MASK2FORMER

We fine-tuned Mask2Former using two approaches. The first involved writing our own code, which was difficult and required a lot of debugging to handle data and model setup. The second approach used Hugging Face's code.

Experimental Setup

- Fine-tuned on our custom dataset from 'Kitti Mots' with COCO-style annotations (train/val splits).
- Used pre-trained Mask2FormerForUniversalSegmentation, fine-tuned with AdamW optimizer (lr=5e-4, weight_decay=1e-4) for 10 epochs, monitoring loss.
- Finetuned the whole model without freezing with and without data augmentation.
- After each epoch, validated the model by calculating average loss and generating instance segmentation predictions, evaluated with COCO metrics.
- Used the fine-tuned model for generating segmentation masks, stored results in JSON format, and evaluated performance using COCO metrics.

BEST EXPERIMENTS AND RESULTS

		mAP@50-95 Person	mAP@50-95 Car
No Freezing AdamW = 5e-4 weight decay = 1e-4	No data augmentation	bbox: 0.19 segm: 0.17	bbox: 0.3 segm:0.28
	A.Compose([A.Perspective(p=0.1), A.RandomBrightnessContrast(p=0.4), A.HueSaturationValue(p=0.1), A.GaussianBlur(p=0.5),])	bbox: 0.02 segm: 0.01	bbox: 0.05 segm: 0.024

FINETUNED RESULTS

- We suspect that something went wrong with the augmentations, as the results didn't seem to make much sense and likely affected the model's performance.
- The fine-tuning didn't show major improvements, possibly because we only trained for 10 epochs, which may not have been enough for significant changes.
- Results were contradictory, we did notice some improvement in detecting people and cars that are further away, indicating the fine-tuning had a slight positive effect in certain cases. But in other cases it was performing worse and missing some significant cars and detecting objects as people so overall it wasn't improving.
- In other words the finetuned model was detecting more false positives compared to True positives.

FINETUNE FOR A DOMAIN SHIFT

Paper link

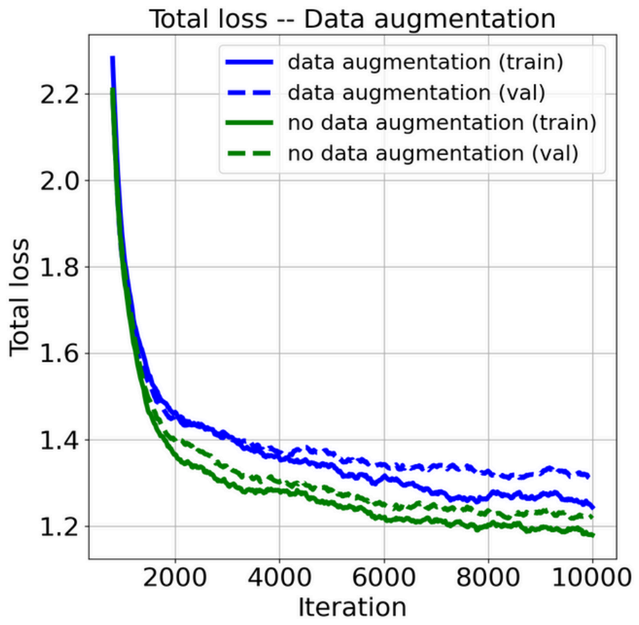
Github link

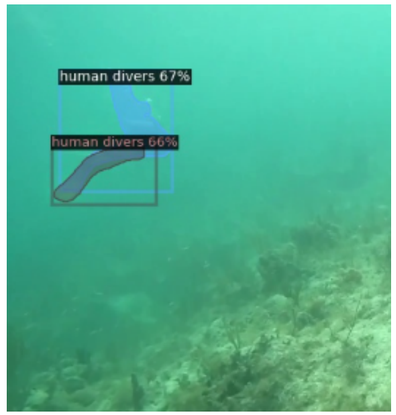
First general Underwater Image Instance Segmentation (UIIS) dataset containing 4,628 images for 7 categories with pixel-level annotations for underwater instance segmentation task.

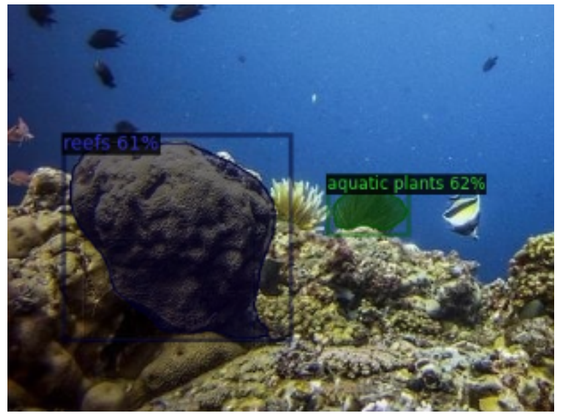
Train	Val	Test
2937 images	1000 images	691 images

	instances	mAP50-95		instances	mAP50-95
fish	2019	30.77	wrecks/ ruins	37	5.05
reefs	1232	12.27	human divers	97	32.67
aquatic plants	201	0.5	robots	17	0.0
sea-floor	145	0.0			

Different learning rates and optimizers were tested, with SGD and a learning rate of 0.0001, running for 10000 iterations, yielding the best results. Various combinations of augmentations were attempted; however, as seen in the graph comparing the loss, the model without augmentation exhibited less overfitting, achieving mAP50 of 20.24. Therefore, the model without augmentation was chosen.







The model performs well detecting certain objects, as seen with the accurate detections of reefs, aquatic plants and human drivers, is also produce many flase positives specially when the object that has to detect are too samll. For example, in the image, it mistakenly classifies fish as reefs and identifies one human dirvers as two. This model could be improved by increasing the number of images labeled for the classes that have less instances, balancing the classes.

ANALYSIS BETWEEN MODELS

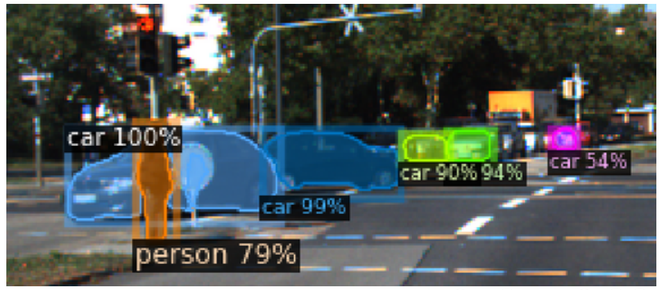
YOLOV11

MASK R-CNN

MASK2FORMER

PRE-TRAINED

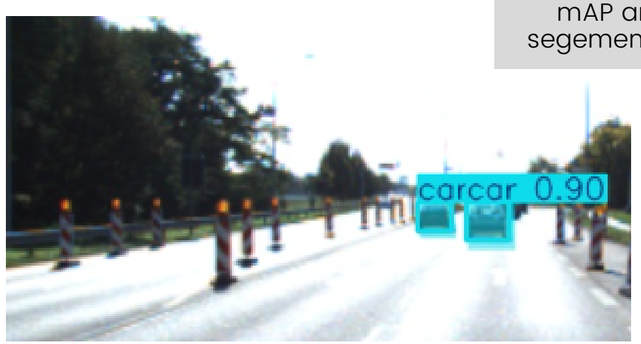
person mAP50-95=0.42
car mAP50-95=0.55



person mAP50-95=0.39
car mAP50-95=0.67

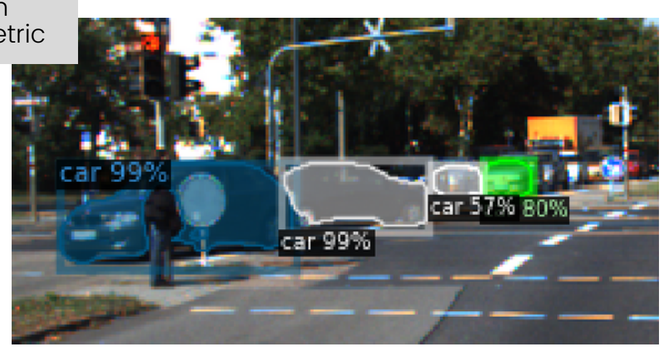
person mAP50-95=0.18
car mAP50-95=0.30

FINETUNED



person mAP50-95=0.43
car mAP50-95=0.57

mAP are all from segementation metric



person mAP50-95=0.38
car mAP50-95=0.68

person mAP50-95=0.17
car mAP50-95=0.28

- In the pre-trained and finetuned stages, Mask R-CNN is the best model in cars segmentation, achieving high mAP scores, but struggle with person detection,which is detected slightly better by YOLO, showing a slight improvement in detection after fine-tuning, particularly in some specific examples.
- After the fine-tuning process, no significant improvements were observed in Mask R-CNN. In fact, in some detections, such as the one in the image, the performance worsened, with the 'person' no longer being detected and confidence over cars being decreased.
- In the case of Mask2Former, no significant improvements were observed. While there were some instances of better performance, such as in the image where it detects many more people, in other cases, the results worsened by detecting people where there are none. Additionally, it is the model with the longest inference time.

INFERENCE TIME	15.2ms	42 ms	4.5 s

FINETUNE

SUMMARY

- We performed fine-tuning for all three proposed models, with YOLO being the model that showed improved results when comparing the fine-tuned model(mAP50-95=0.43 for person and for car mAP50-95=0.57) and the pre-trained one(mAP50-95=0.42 for person and for car mAP50-95=0.55).
- Neither Mask R-CNN nor Mask2Former showed significant improvements after fine-tuning. In both cases, there were examples where objects that were previously missed were correctly detected, but also instances where objects that were originally well-detected were mistakenly identified.
- YOLOv11 achieved the fastest inference time, making it the most efficient choice in our case.

DOMAIN SHIFT

First general Underwater Image Instance Segmentation (UIIS) dataset containing 4,628 images for 7 categories with pixel-level annotations for underwater instance segmentation task.

Various augmentation strategies were tested, and the non-augmented model achieved a higher mAP of 20, and exhibited less overfitting.

The model performs well detecting certain objects, as seen with the accurate detections of reefs, aquatic plants and human drivers, is also produce many flase positives specially when the object that has to detect are too samll. For example, in the image, it mistakenly classifies fish as reefs and identifies one human dirvers as two.

This model could be improved by increasing the number images labeled for the classes that have less instances, balancing the classes.