UNIVERSITÉ DE TOULON

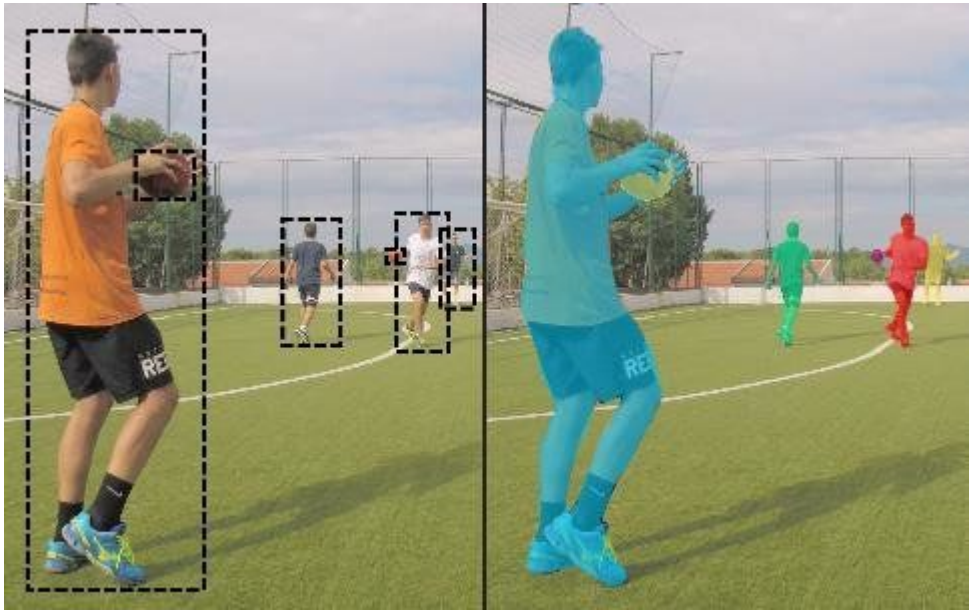
MIR MARINE & MARITIME INTELLIGENT ROBOTICS

Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and A Large-scale Dataset

Muhammad Pramudya, Hyejoo Kwon, Mateus Pedrosa, Youssief Morsy

AI Project 1: Under the Supervision of Prof. Marxer

What is image segmentation ?



Precise object localization

better size estimation

Enhanced object recognition

Problem that is addressed by USIS

the lack of open source dataset available for labelling

Meta

ultralytics

pre-existing popular model does not include underwater

underwater images often suffer from color distortion

Paper Contributions

- USIS10K dataset with annotated underwater images
- USIS-SAM model to improve the SIS task in underwater scenes.

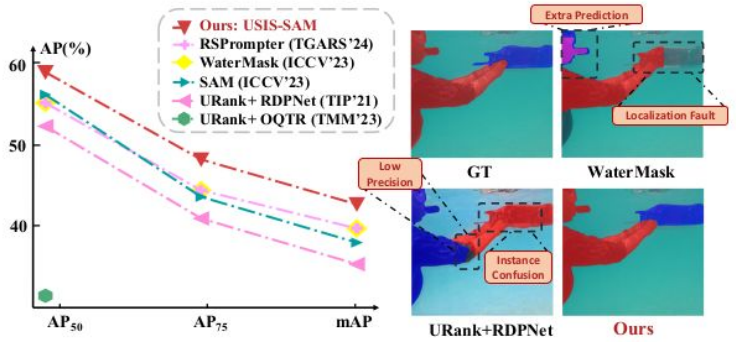


Figure 1. A simple comparison of USIS-SAM and other state-of-the-art methods trained on the USIS10K dataset. Different colors represent different salient instances. URank represents the underwater image enhancement method in UnderwaterRanker (Guo et al., 2023).

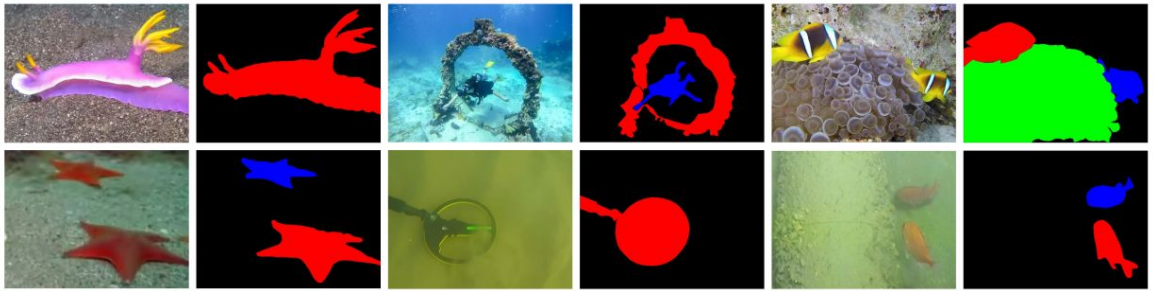


Figure 2. Examples of annotations for various salient instances in USIS10K. The image on the left is the original image and the right is the annotation mask, different colors represent different salient instances. More dataset showings can be found in Appendix A.4.

USIS10K Dataset

- First large-scale dataset for USIS task
- 10,632 images
- Pixel-level annotations of 7 categories
- Images from the internet and open source underwater datasets from different domains
- Class-Agnostic and Multi-Class labels
- Training, validation and test sets (7:1.5:1.5)

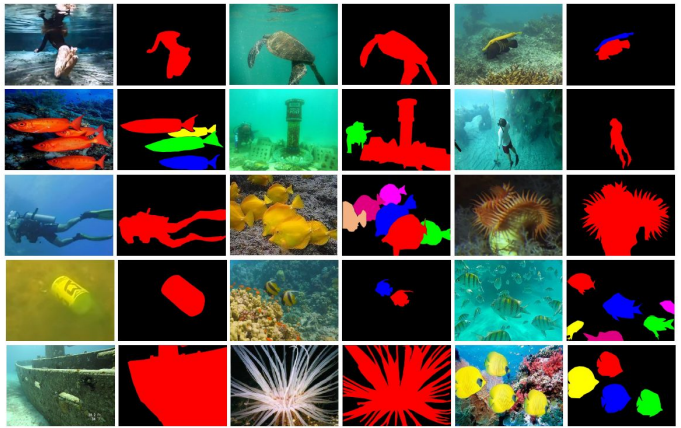


Figure 9. **More visual samples of annotated images with pixel-level salient instance segmentation in USIS10K.** The image on the left is the original image and the right is the annotation mask, different colors represent different salient instances.

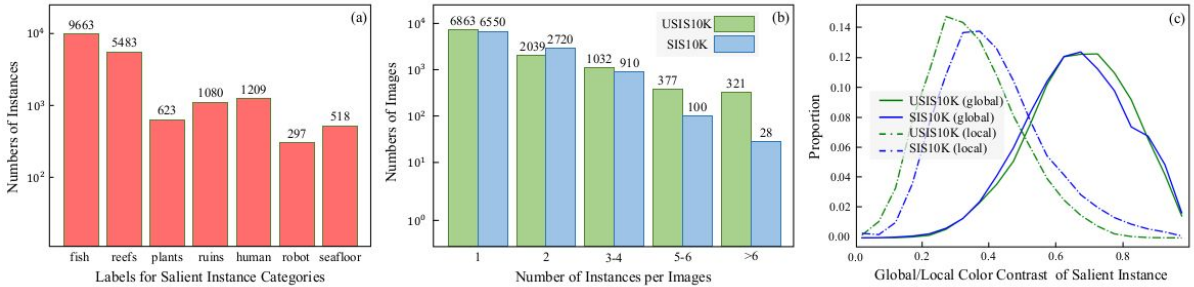
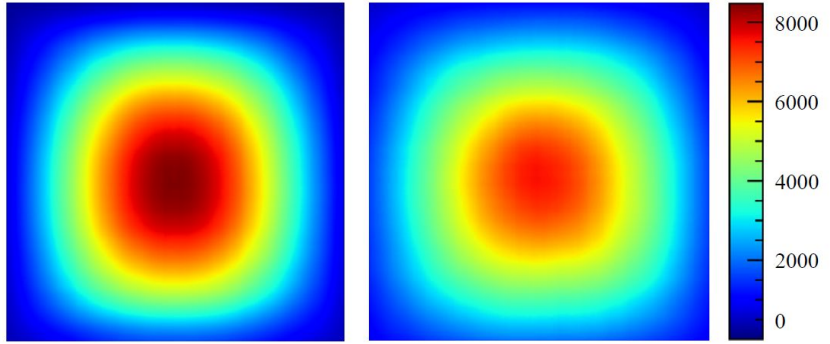


Figure 3. Essential characteristics of the USIS10K dataset. (a) The number of salient instances per category in the USIS10K dataset. (b) Distribution of the number of salient instances per image in the USIS10K dataset. (c) Comparison of USIS10K and SIS10K in global color contrast and local color contrast.

Dataset



(a) SIS10K (b) USIS10K

Figure 4. A set of salient maps from our dataset and SIS10K.

Table 5. The category labels in USIS10K dataset with their definitions.

Category	Descriptions
Fish	Underwater vertebrates, e.g., fish, turtles
Reefs	Underwater invertebrates and coral reefs
Aquatic plants	Aquatic plants and flora
Wrecks/ruins	Wrecks, ruins and damaged artifacts
Human divers	Human divers and their equipment
Robots	Underwater robots like AUV, ROV
Sea-floor	Rocks and reefs on the seafloor

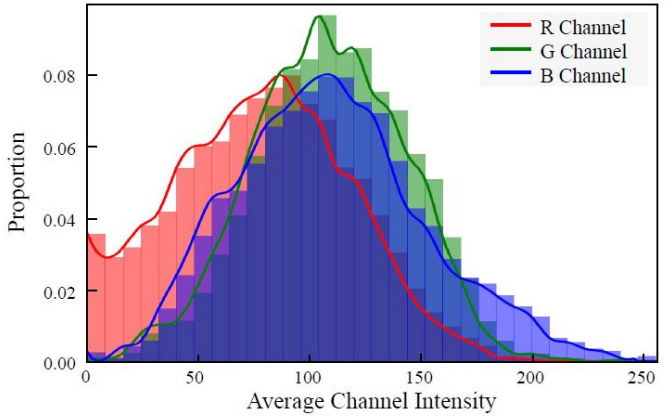


Figure 5. Average channel intensity in USIS10K with proportion.

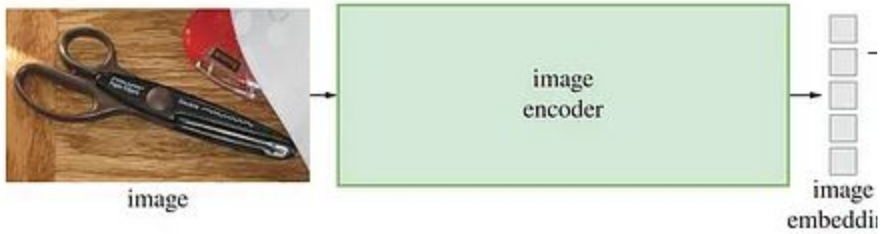
SAM

- Image encoder (masked autoencoder) to extract the image embedding,
- Prompt encoder that takes in different types of prompt
- Mask decoder to build a mask.

https://medium.com/@utkarsh135/segment-anything-model-sam-explained-2900743cb61e

SAM

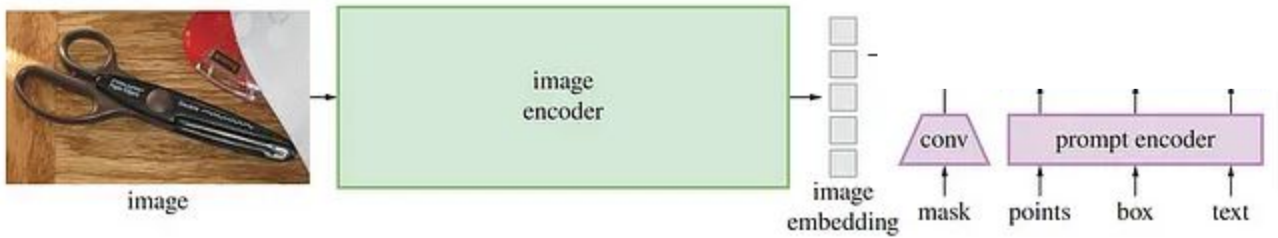
- Image encoder (masked autoencoder) to extract the image embedding,
- Prompt encoder that takes in different types of prompt
- Mask decoder to build a mask.



https://medium.com/@utkarsh135/segment-anything-model-sam-explained-2900743cb61e

SAM

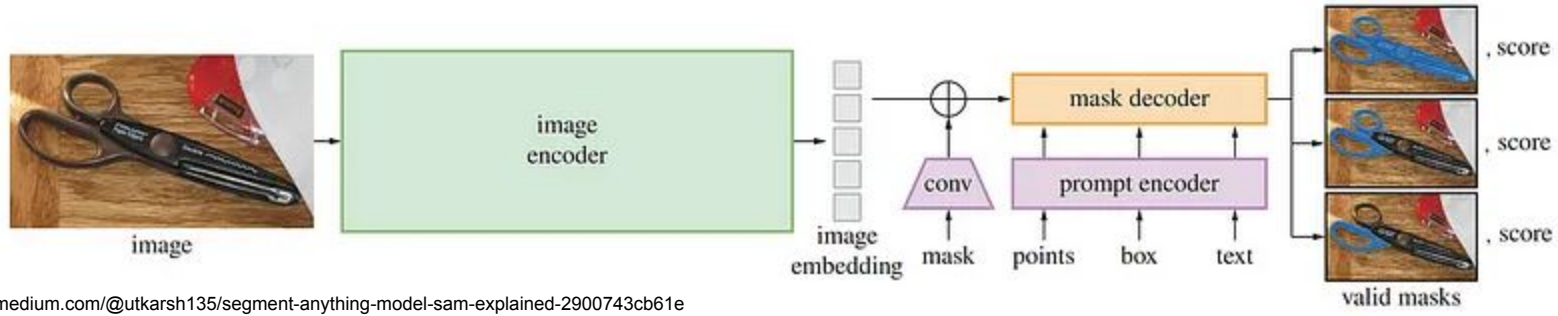
- Image encoder (masked autoencoder) to extract the image embedding,
- Prompt encoder that takes in different types of prompt
- Mask decoder to build a mask.



https://medium.com/@utkarsh135/segment-anything-model-sam-explained-2900743cb61e

SAM

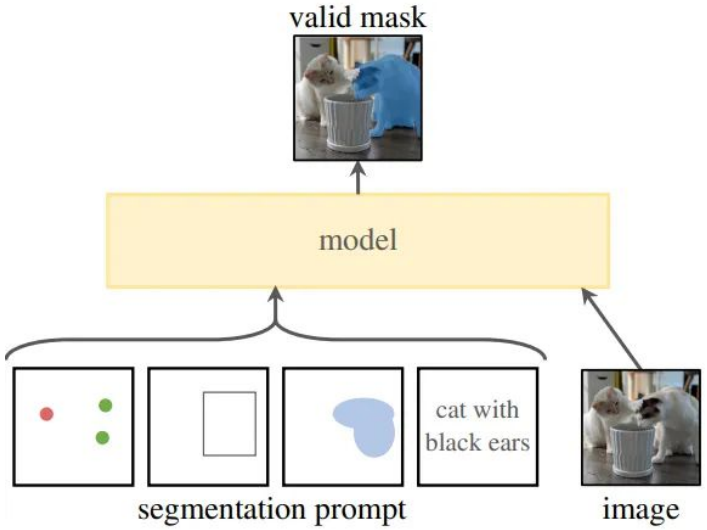
- Image encoder (masked autoencoder) to extract the image embedding,
- Prompt encoder that takes in different types of prompt
- Mask decoder to build a mask.



https://medium.com/@utkarsh135/segment-anything-model-sam-explained-2900743cb61e

SAM

- Image encoder (masked autoencoder) to extract the image embedding,
- Prompt encoder that takes in different types of prompt
- Mask decoder to build a mask.



https://medium.com/@utkarsh135/segment-anything-model-sam-explained-2900743cb61e

USIS-SAM

- Underwater Adaptive ViT (UA-ViT)
  - Adapter
  - Channel adapter
- A Salient Feature Prompt Generator (SFPG)
  - SFFM
  - RPN

Model Architecture

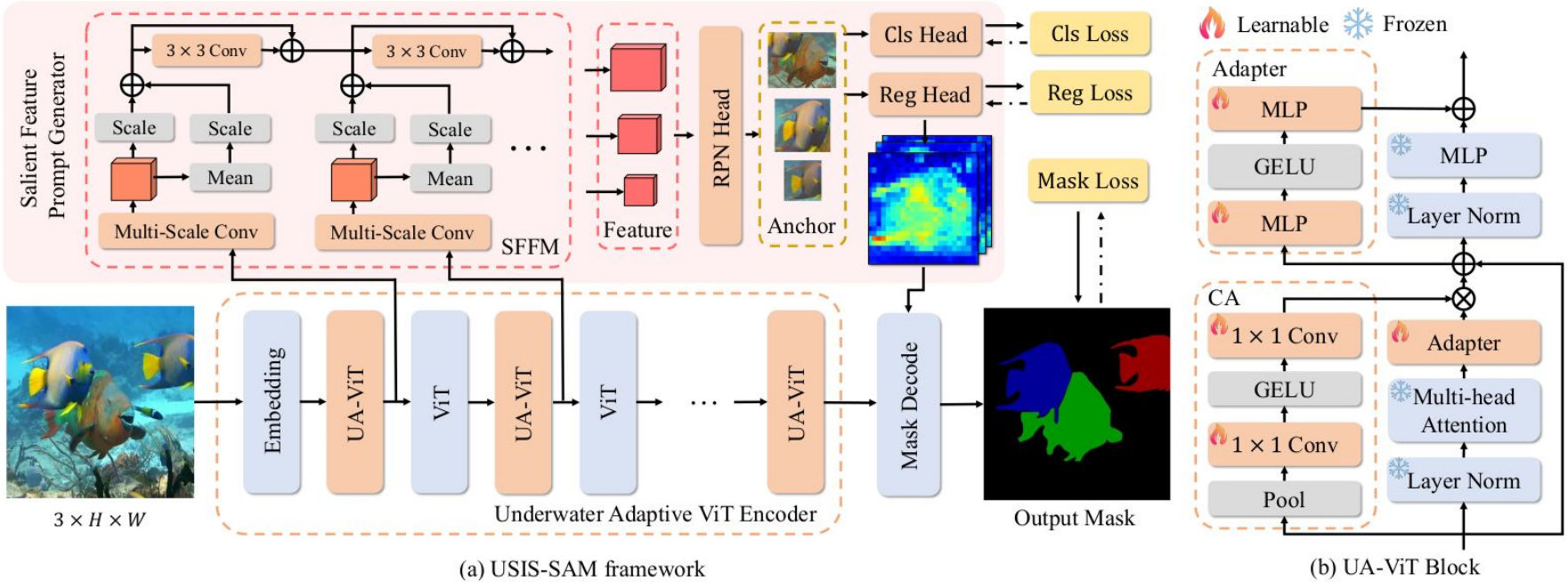
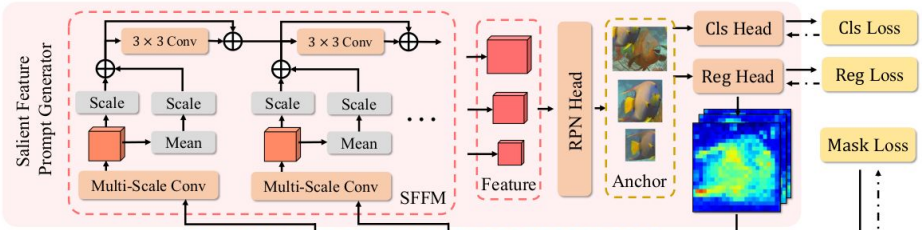


Figure 6. (a) USIS-SAM framework. The USIS-SAM framework modifies the SAM by adding the Underwater Adaptive ViT Encoder (in Section 4.1) and the Salient Feature Prompt Generator (in Section 4.2). (b) The structure of UA-ViT. In the figure, SFFM stands for Salient Feature Fusion Module, CA stands for Channel Adapter.

Salient Feature Prompt Generator (SFPG)

- Salient Feature Fusion Module (SFFM)
- Region Proposal Network (RPN)



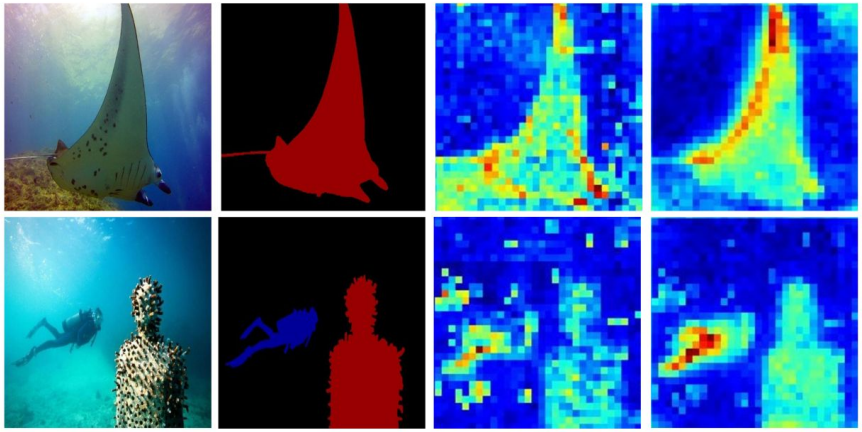
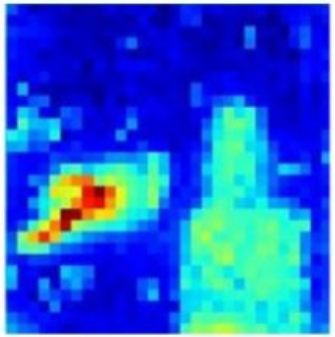


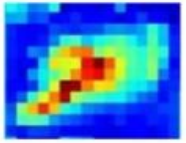
Figure 7. Visualize features generated by the SFPG. The SFPG can aggregate features on salient instances.

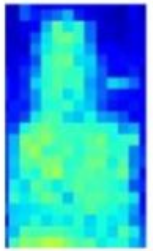
SFPG

SFFM



RPN





Training details

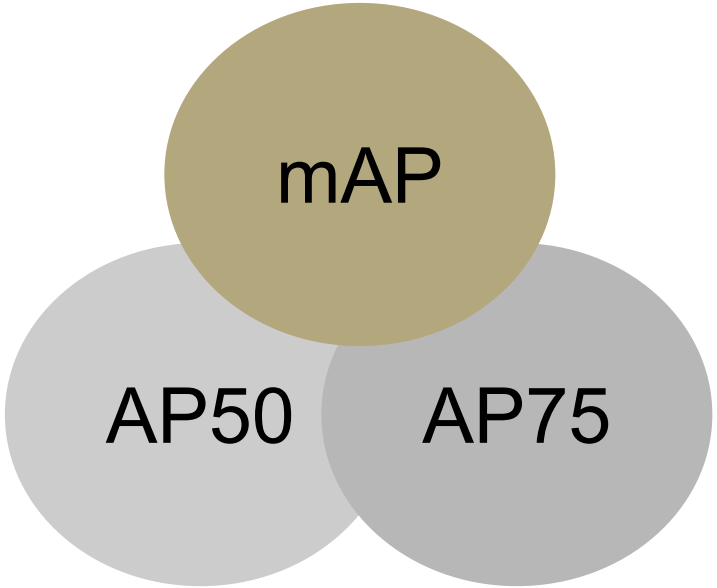
Training Setup:

- **Epochs**: 24 epochs.
- **Hardware**: 6 NVIDIA 3090 GPUs.
- **Optimizer**: AdamW.

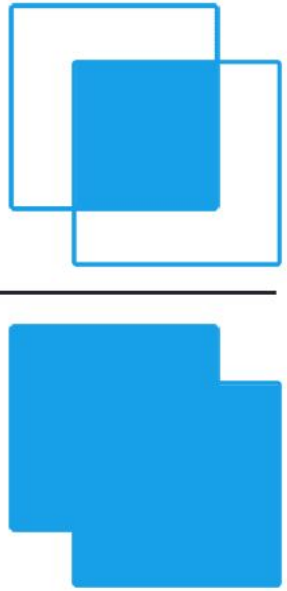
Hyperparameters:

- **Learning Rate**: 1e-4.
- **Weight Decay**: 1e-3.

How to Measure Performance

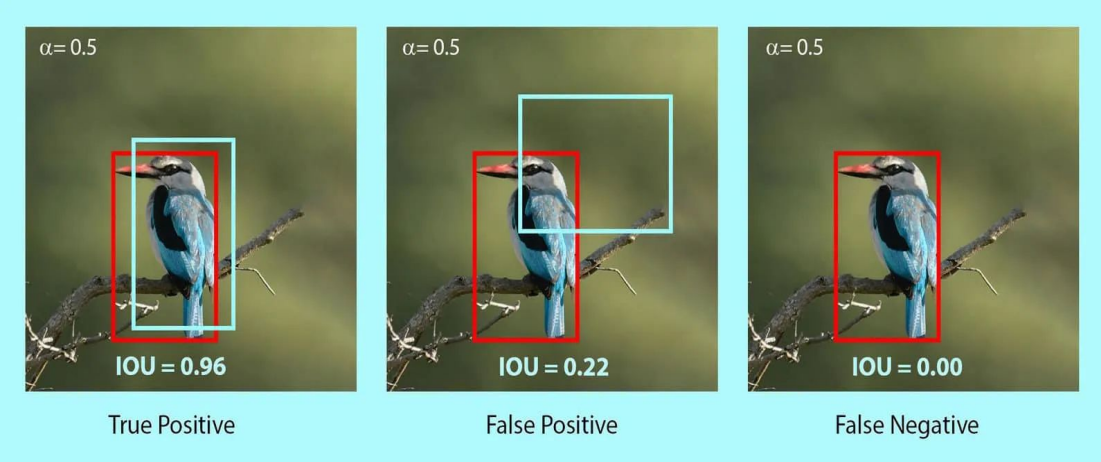


$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



https://learnopencv.com/intersection-over-union-iou-in-object-detection-and-segmentation/

How to Measure Performance

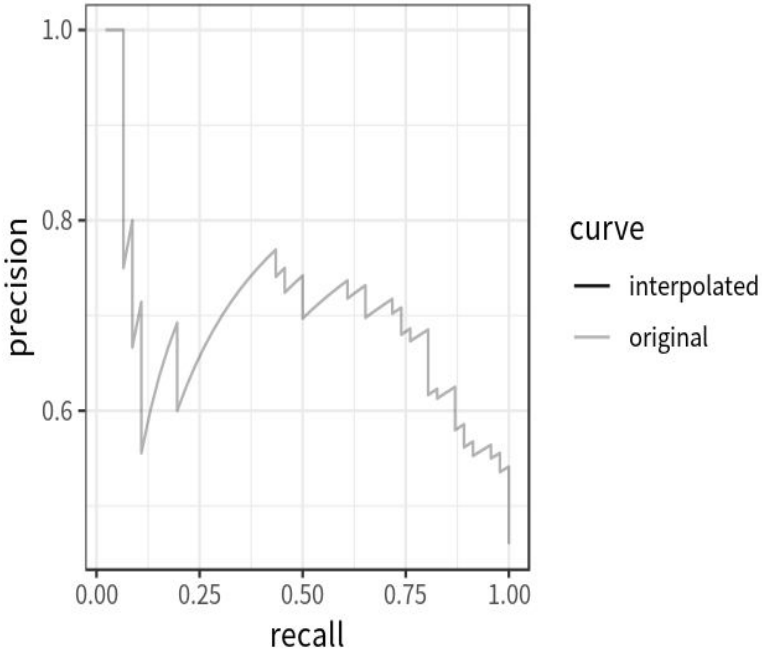


$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$AP = \int_0^1 P(r)dr$$

→ **AP50 AP75 mAP**



https://learnopencv.com/intersection-over-union-iou-in-object-detection-and-segmentation/

Model Evaluation

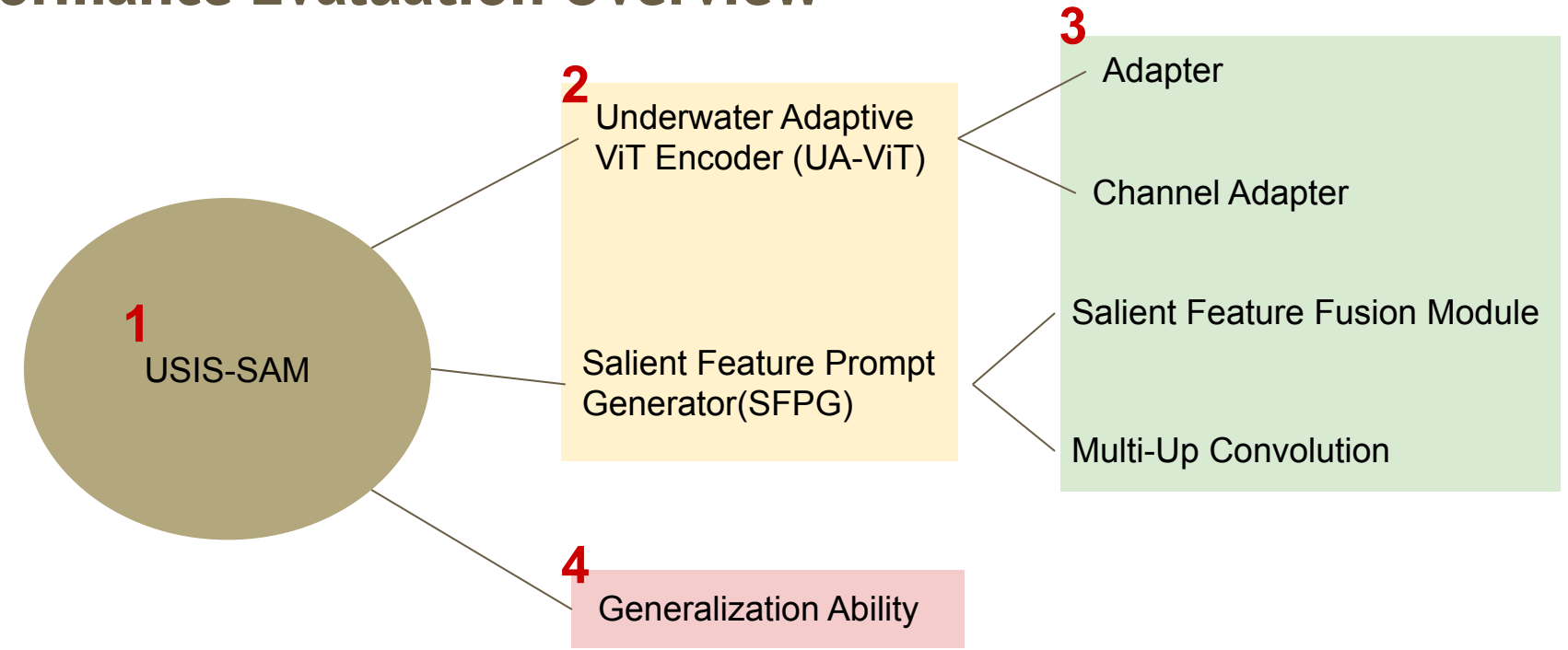
All backbones & hyperparameters are the same as original paper

Method	Epoch	Backbone	Class-Agnostic			Multi-Class		
			mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
S4Net (Fan et al., 2019)	60	ResNet-50	32.8	64.1	27.3	23.9	43.5	24.4
RDPNet (Wu et al., 2021)	50	ResNet-50	53.8	77.8	61.9	37.9	55.3	42.7
RDPNet (Wu et al., 2021)	50	ResNet-101	54.7	78.3	63.0	39.3	55.9	45.4
OQTR (Pei et al., 2023)	120	ResNet-50	56.6	79.3	62.6	19.7	30.6	21.9
URank+RDPNet (Wu et al., 2021)	50	ResNet-101	52.0	80.7	62.0	35.9	52.5	41.4
URank+OQTR (Pei et al., 2023)	120	ResNet-50	49.3	74.3	56.2	20.8	32.1	23.3
WaterMask (Lian et al., 2023)	36	ResNet-50	58.3	80.2	66.5	37.7	54.0	42.5
WaterMask (Lian et al., 2023)	36	ResNet-101	59.0	80.6	67.2	38.7	54.9	43.2
SAM+BBox (Kirillov et al., 2023)	24	ViT-H	45.9	65.9	52.1	26.4	38.9	29.0
SAM+Mask (Kirillov et al., 2023)	24	ViT-H	55.1	80.2	62.8	38.5	56.3	44.0
RSPrompter (Chen et al., 2023a)	24	ViT-H	58.2	79.9	65.9	40.2	55.3	44.8
URank+RSPrompter (Chen et al., 2023a)	24	ViT-H	50.6	74.4	56.6	38.7	55.4	43.6
USIS-SAM	24	ViT-H	**59.7**	**81.6**	**67.7**	**43.1**	**59.0**	**48.5**

Class-Agnostic
Detects all objects as a single type, without classifying them

Multi-Class
Detects objects and classifies them into different categories

Performance Evaluation Overview



Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o UA-ViT	41.5 (-1.6)	57.4 (-1.6)	47.0 (-1.5)
replace SFPG	42.2 (-0.9)	58.3 (-0.7)	47.5 (-1.0)

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o Adapter	41.7 (-1.4)	57.3 (-1.7)	47.3 (-1.2)
w/o CA	42.0 (-1.1)	57.7 (-1.3)	47.1 (-1.4)

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o SFFM	42.3 (-0.8)	58.5 (-0.5)	47.2 (-1.3)
w/o Multi-Up	42.5 (-0.6)	58.6 (-0.4)	47.7 (-0.8)

Backbone : A backbone is the main part of a model that extracts features from an image, like shapes, colors, and patterns. It acts as the 'base network' that other parts of the model use to make decisions, such as detecting objects or segmenting image

Methods	Epoch	mAP	AP_{50}	AP_{75}
Mask RCNN (He et al., 2017)	36	23.4	40.9	25.3
Point Rend (Kirillov et al., 2020)	36	25.9	43.4	27.6
QueryInst (Fang et al., 2021)	36	26.0	42.8	27.3
Mask2Former (Cheng et al., 2022)	36	25.7	38.0	27.7
RDPNet (Wu et al., 2021)	50	20.6	38.7	19.4
WaterMask (Lian et al., 2023)	36	27.2	43.7	29.3
USIS-SAM	24	**29.4**	**45.0**	**32.3**

Our Work / Implementation

- USIS-SAM Model Training Results and Inference
  - Train model on the cluster
  - 24 epochs completed in 20 hours
  - Trained on A100:20 GPUs
  - Results comparable to original paper
  - we needed to reduce the batch size to be able to train the model
- Compare the Performance of other Segmentation Models Yolo.

Comparison : YoloV11 Segmentation

pretrained weights fine-tuned with our datasets

Models

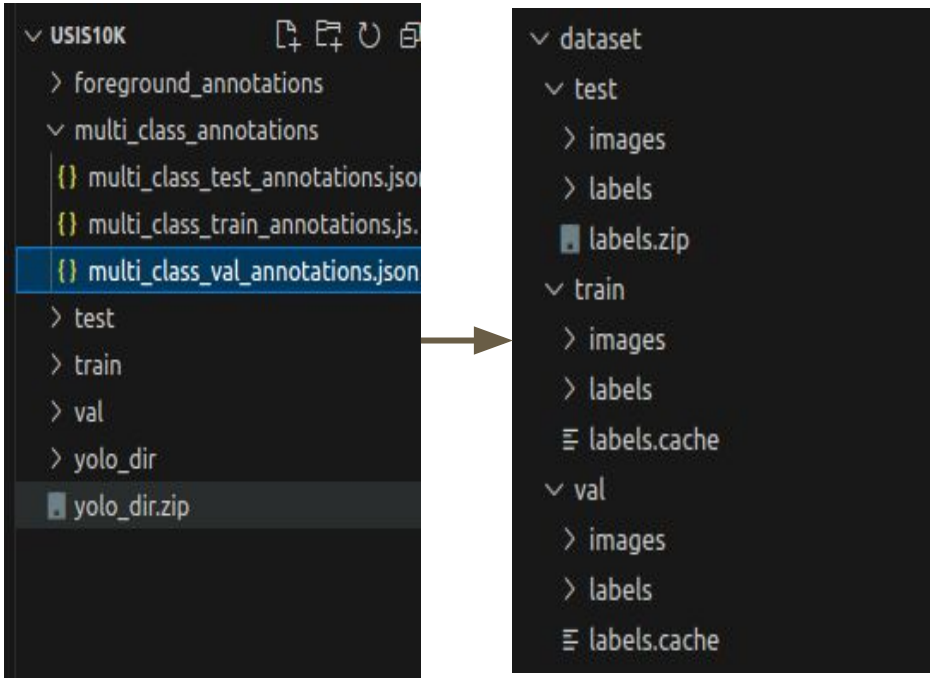
YOLO11 pretrained Segment models are shown here. Detect, Segment and Pose models are pretrained on the COCO dataset, while Classify models are pretrained on the ImageNet dataset.

Models download automatically from the latest Ultralytics release on first use.

Model	size (pixels)	mAP^{box} 50-95	mAP^{mask} 50-95	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n-seg	640	38.9	32.0	65.9 ± 1.1	1.8 ± 0.0	2.9	10.4
YOLO11s-seg	640	46.6	37.8	117.6 ± 4.9	2.9 ± 0.0	10.1	35.5
YOLO11m-seg	640	51.5	41.5	281.6 ± 1.2	6.3 ± 0.1	22.4	123.3
YOLO11l-seg	640	53.4	42.9	344.2 ± 3.2	7.8 ± 0.2	27.6	142.2
YOLO11x-seg	640	54.7	43.8	664.5 ± 3.2	15.8 ± 0.7	62.1	319.0

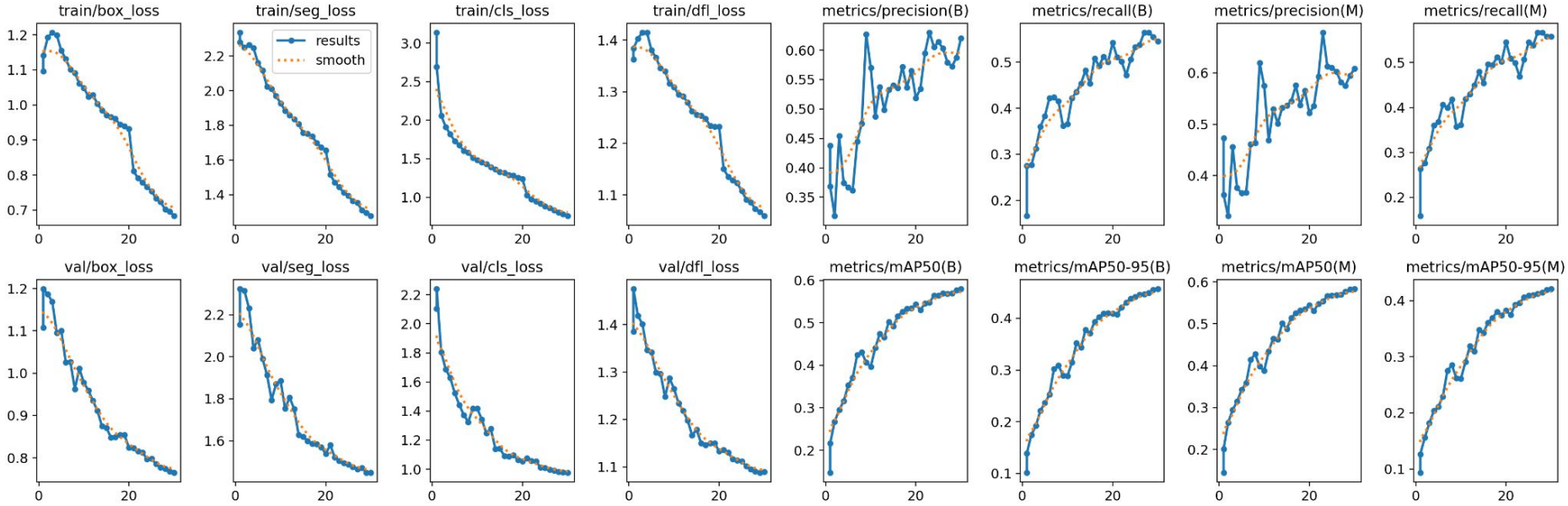
SCAN ME

Dataset and training configuration

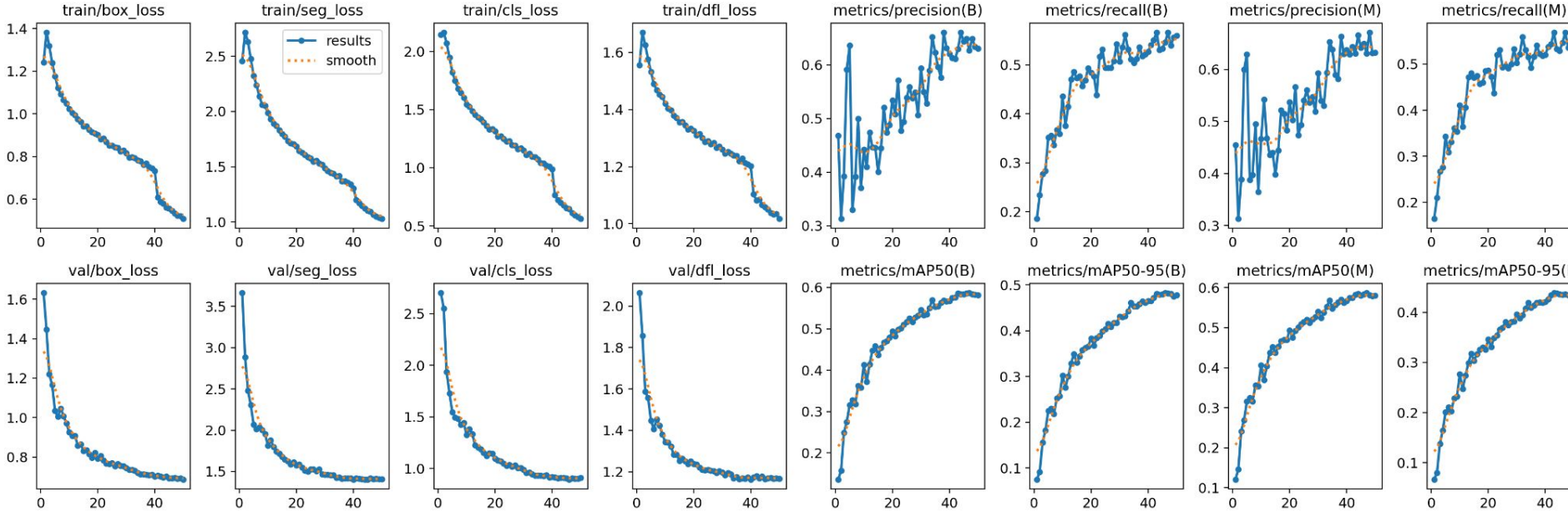


Parameter	Details
Dataset Size	~10,000 Images
Training Set	7,500 Images
Validation Set	1,000 Images
Test Set	1,500 Images
Batch Size	16
Training Time	~4 Hours
Hardware Used	NVIDIA A100 (20GB VRAM)

Training results YoloV11 Nano

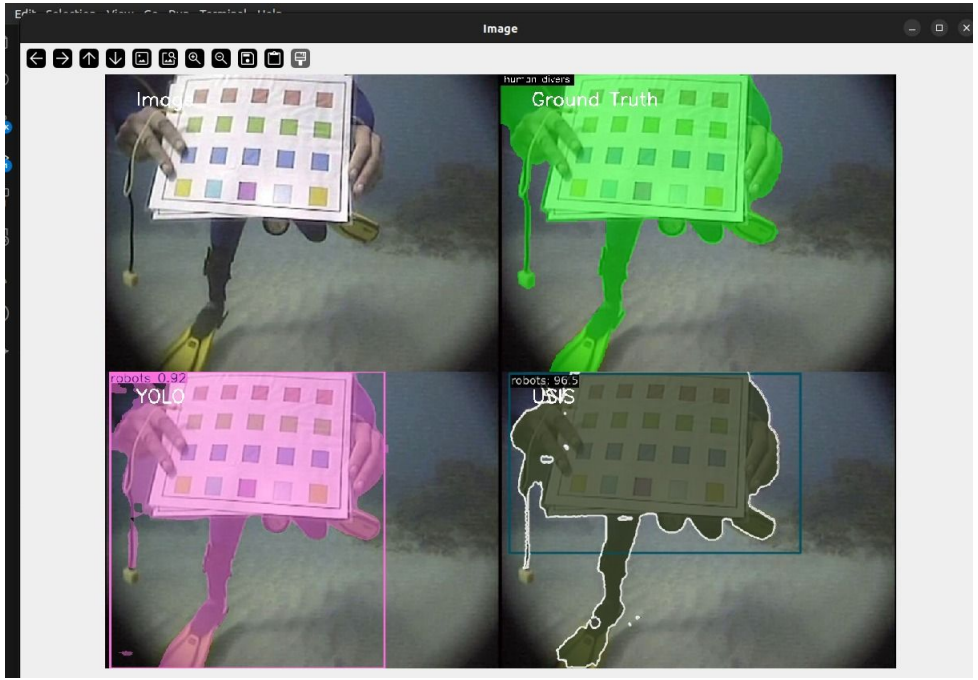


Training results XL



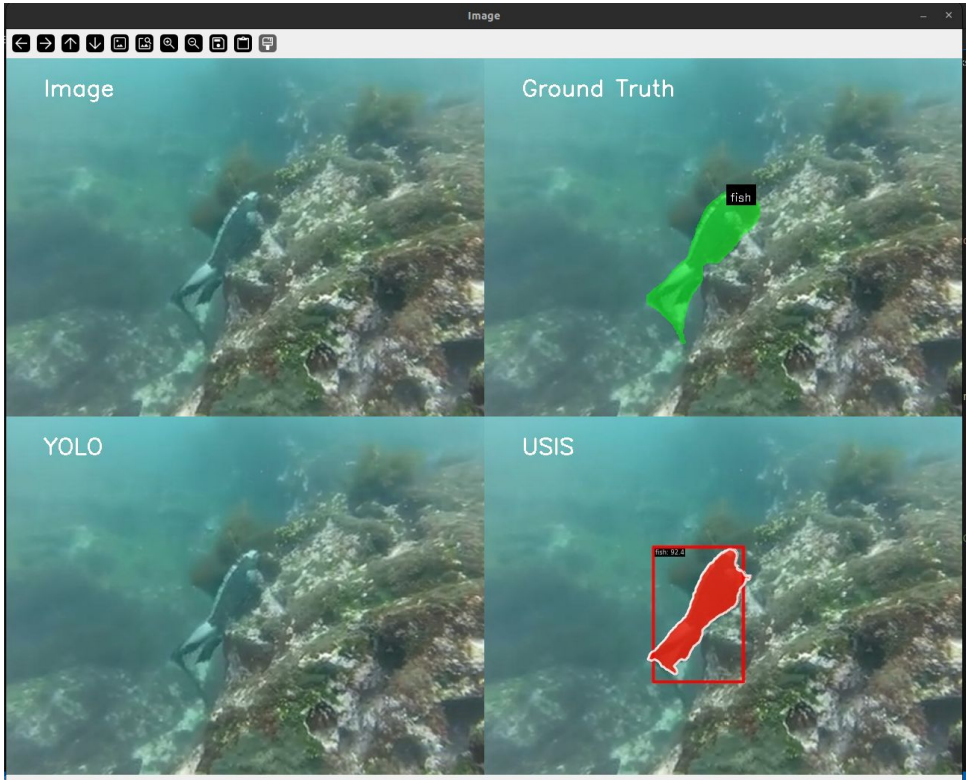
Results Comparison

quantitative comparison			
	epoch	mAP	mAp50
USIS-SAM	24	43.1	59.0
YoloV11X	24	51.05	51.05
YoloV11n	24	54.79	43.07
YoloV11n	50	57.94	45.63
YoloV11X	50	58.13	57.79

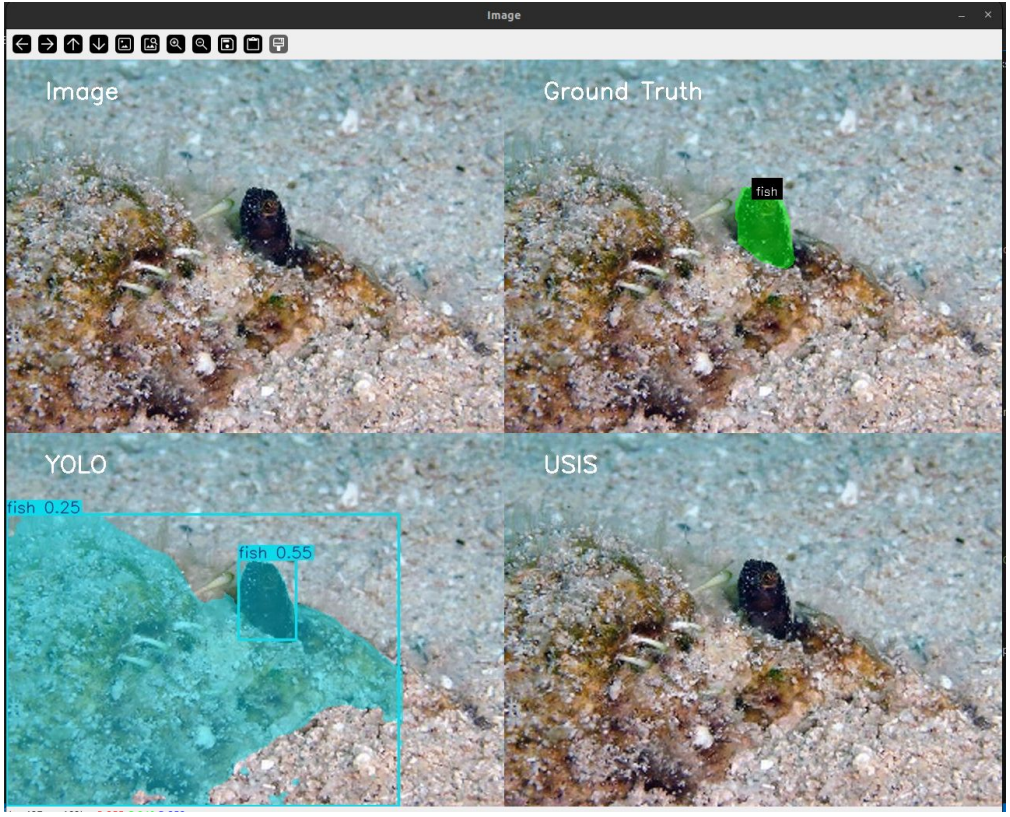


mAP : Mean Average Precision
mAP50 : for IoU >0.5

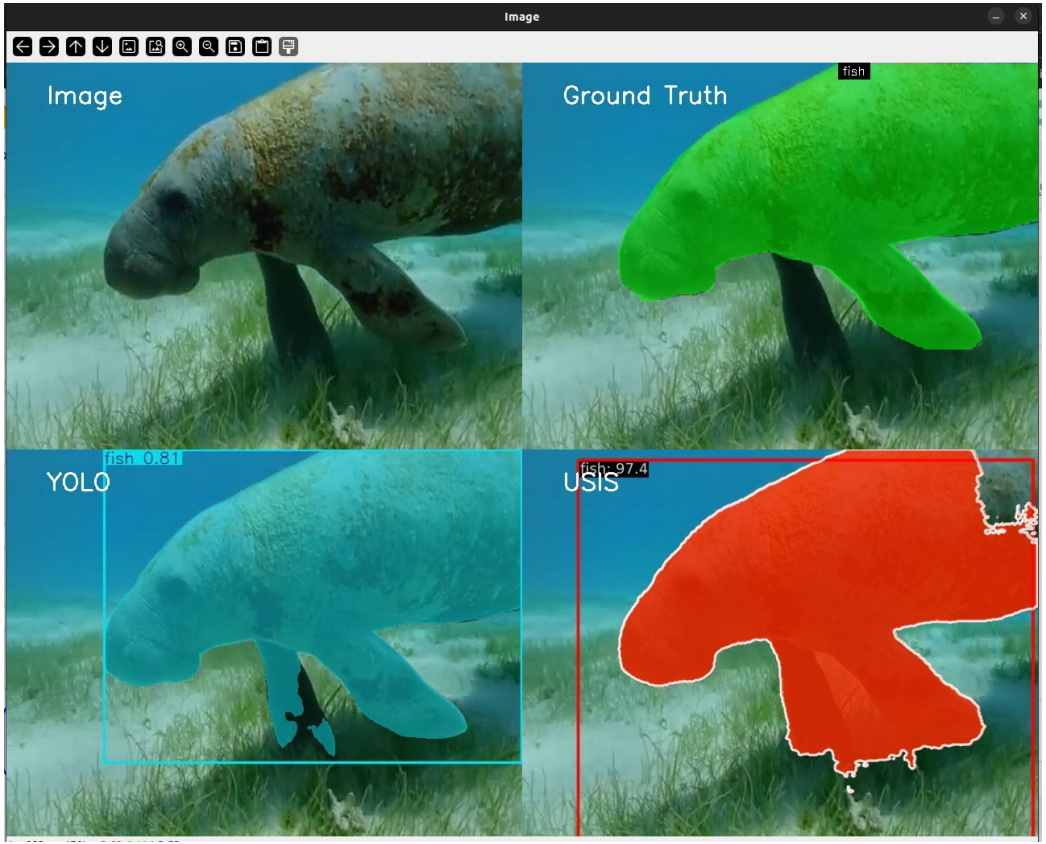
Results: USIS SUCCESS YOLO FAIL



Results: Results: USIS FAIL YOLO SEMI-FAIL



Results



Results

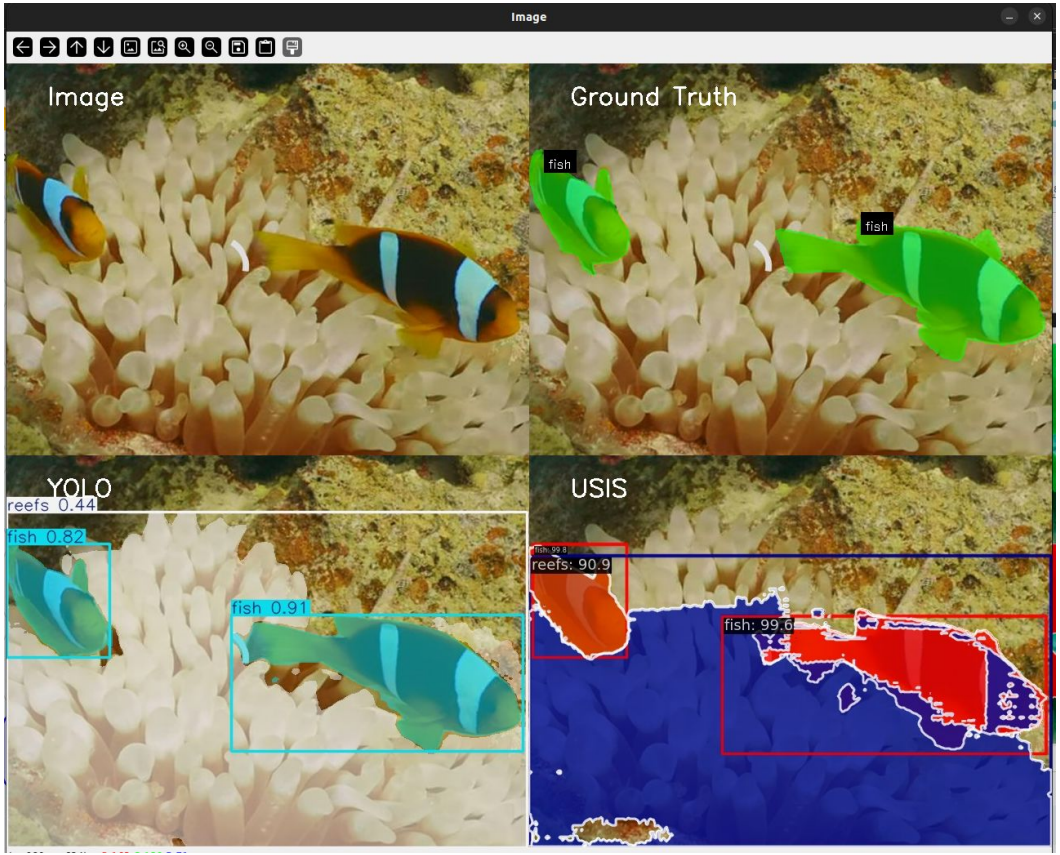
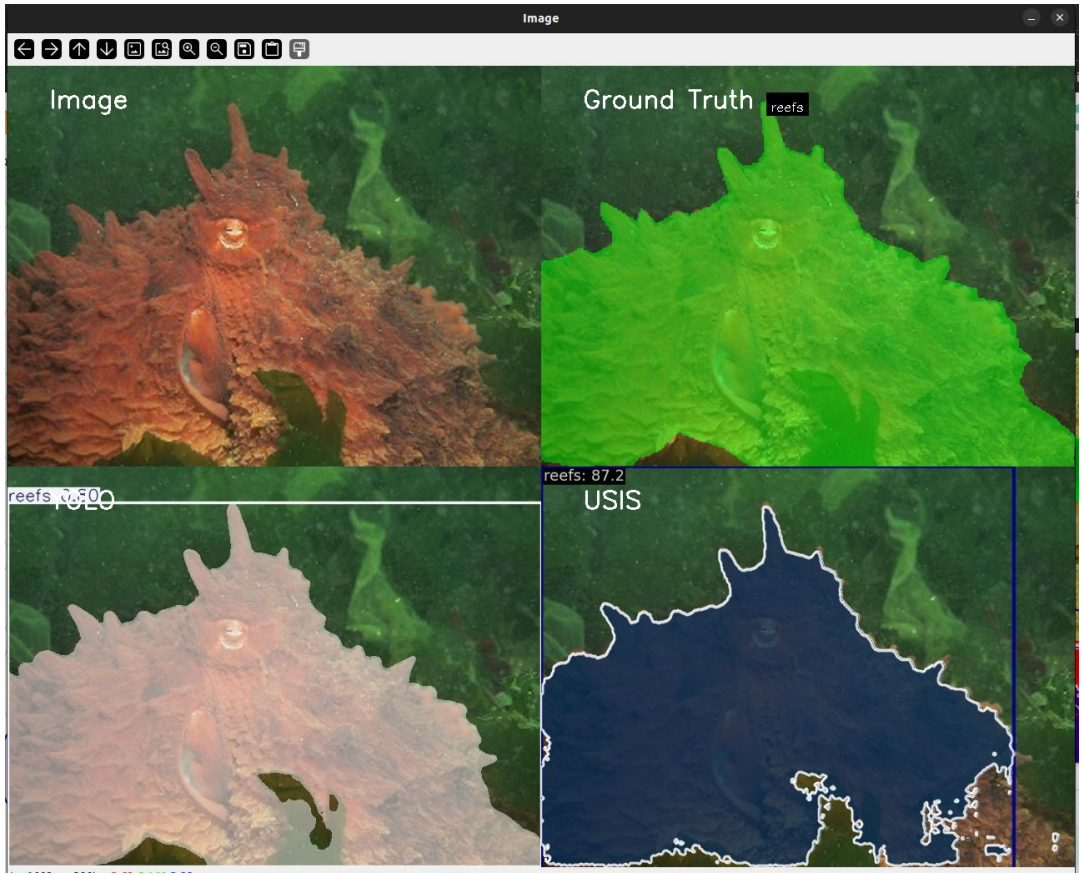


Image
Image
Ground Truth
fish
fish
YOLO
reefs 0.44
fish 0.82
fish 0.91
USIS
reefs: 90.9
fish: 99.6

Results



Results

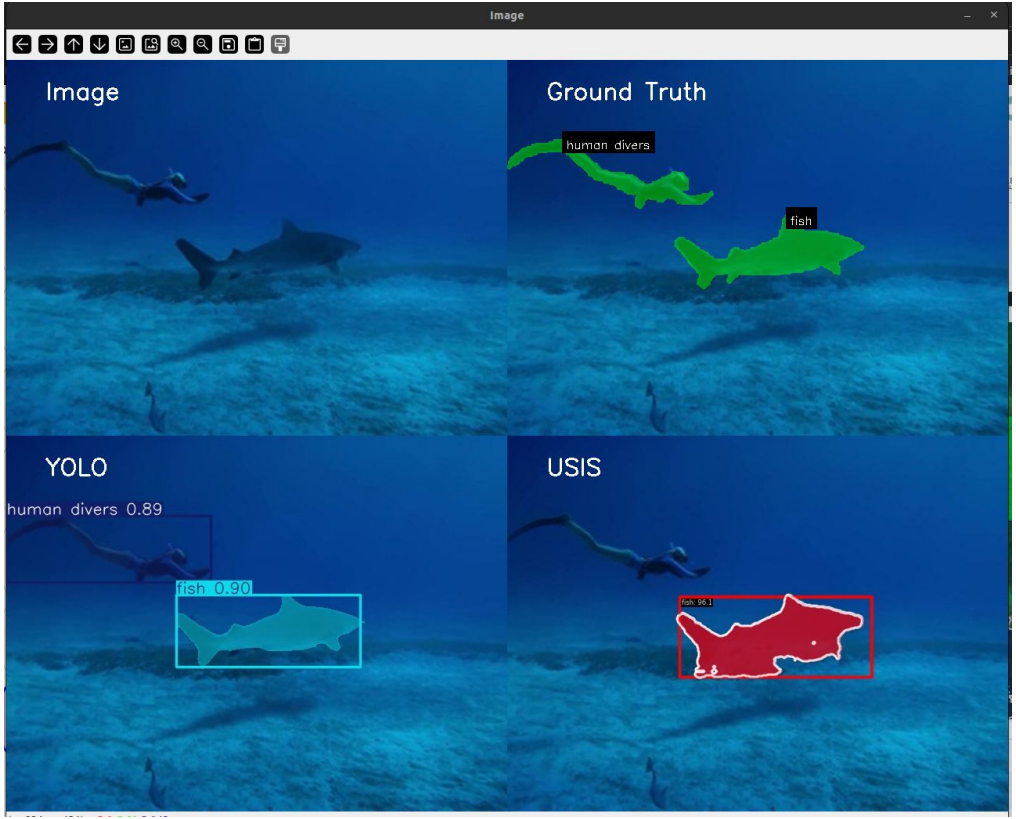


Image
Image
Ground Truth
human divers
fish
YOLO
human divers 0.89
fish 0.90
USIS

Thank you