FishID

Automating the Detection & Classification of Fish from Underwater Video

Objective

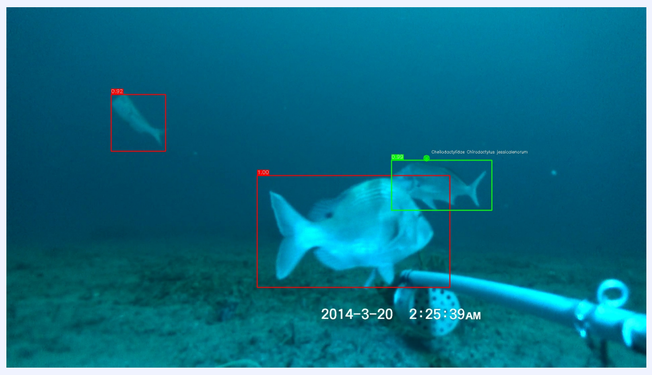
- Marine biologists, fisheries managers and ecologists use underwater cameras to monitor fish population size and composition
- Current monitoring techniques require manually annotating fish in underwater footage
- Using footage collected by the South African Institute for Aquatic Biodiversity (SAIAB), we investigated DL approaches to automating this labour-intensive process

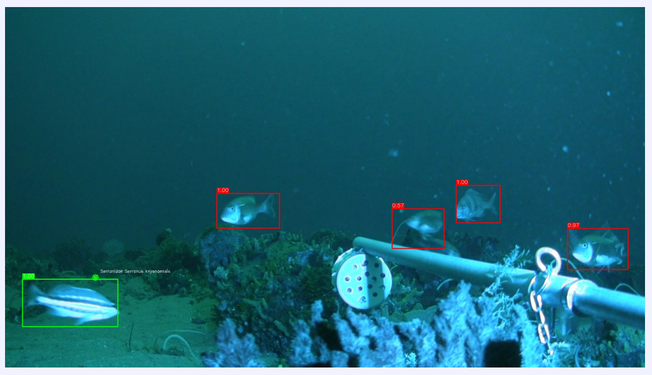
Dataset

- The dataset provided by SAIAB posed challenges for effective application of DL
- The dataset contained several hundred hours of video, however, only a small number of frames were only partially labelled

Preprocessing

- Designed software to perform semi-automated data labelling using an open source fish-detection model.
- Allowed the team to annotate ~60,000 unlabelled fish samples and ~5,000 labelled fish samples





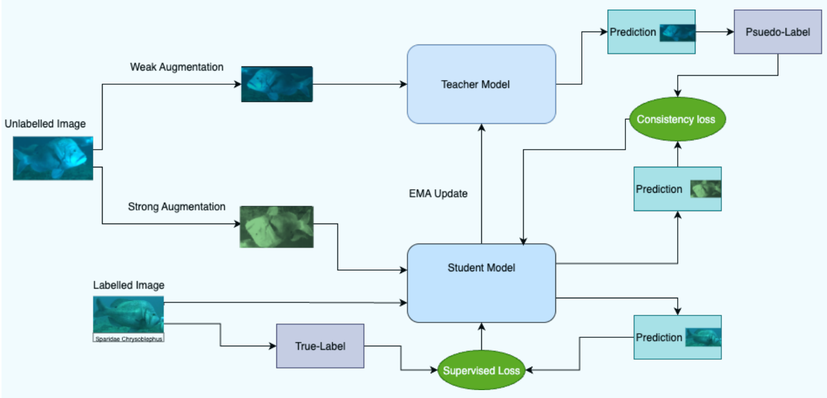
Detector

- Using the unlabelled samples, the team trained a custom YOLOv11 detector that achieved a mAP@50 of **80.2%.**

Classification Approaches

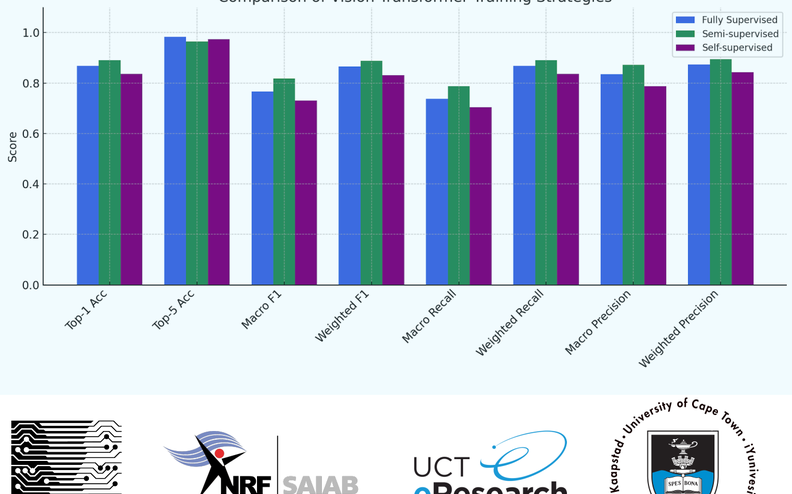
Semi-Supervised Learning with Vision Transformers

- By leveraging both labelled and unlabelled samples, trained a Vision Transformer (ViT) using an Exponential Moving Average (EMA) teacher-student framework
- This enabled the model to learn robust representations despite limited labelled data, addressing one of the major challenges in underwater species recognition



Results

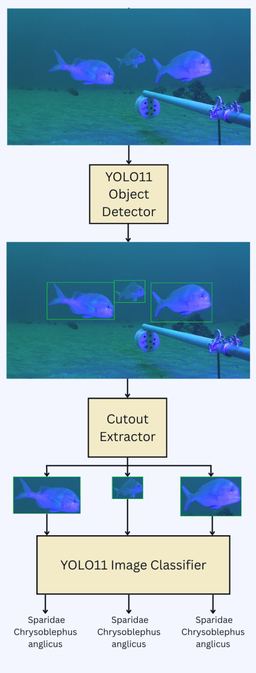
- ViT-EMA achieved highest top-1 accuracy (89.0%), outperforming fully supervised and MAE-pretrained variants
- Improvements also observed in macro and weighted F1-scores, indicating stronger generalisation across underrepresented fish classes
- However, MAE pre-training hindered performance, suggesting poor alignment with the classification task



Object Tracking for Data Augmentation

- Object tracking enabled better utilisation of the given video dataset
- By tracking labelled fish across consecutive video frames, generated a larger, more diverse labelled dataset for training

- Developed a two-stage detection-classification pipeline that identifies fish within video frames and classifies each detection by species.



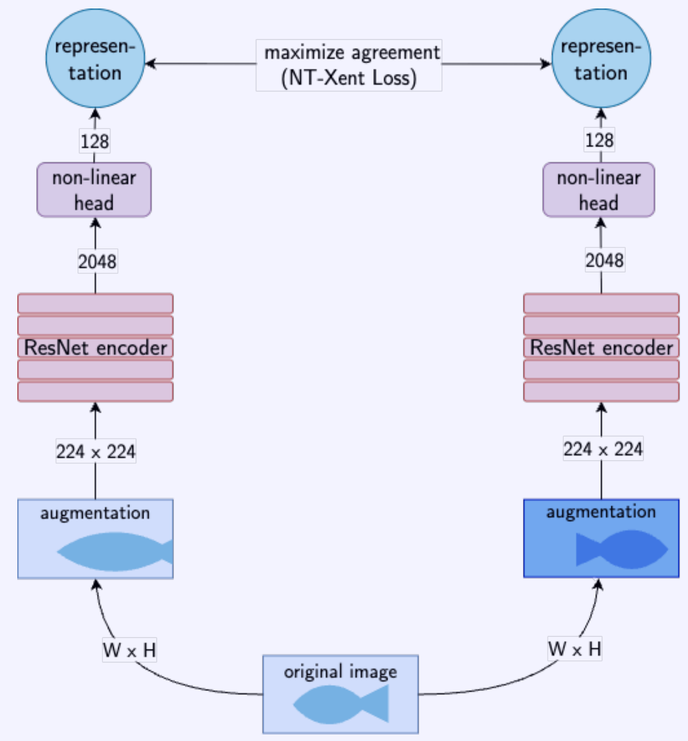
Results

- The YOLOv11 classifier trained on the tracking-augmented dataset achieved 13.4% higher top-1 accuracy compared to the model trained on the original dataset
- The augmented model also showed improvements in both macro and weighted F1 scores
- The model demonstrated better classification performance across all fish classes

Model	Train Data	Top-1	Top-3	Top-5	Macro F1	Weighted F1
ResNet	Original	0.4176	0.6287	0.7454	0.1914	0.4593
ResNet	Augmented	0.4156	0.6673	0.7618	0.2172	0.4651
YOLOv11	Original	0.6577	0.8255	0.8872	0.4725	0.6751
YOLOv11	Augmented	**0.7917**	**0.9036**	**0.9402**	**0.6087**	**0.7999**

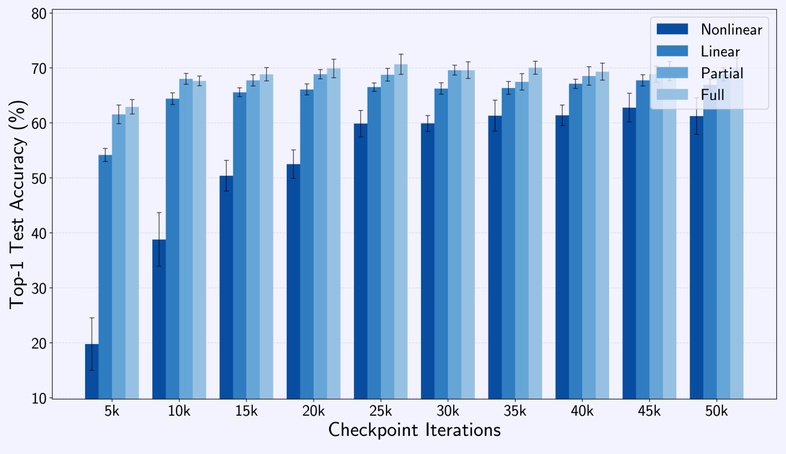
SimCLR Contrastive Learning

- Contrastive Learning is a self-supervised DL technique that allows for unlabelled data to be utilised for effective feature learning
- This allowed for large quantities of unlabelled data could be taken advantage of during training



Results

- Models learnt effective features through SimCLR training, plateauing after 25K steps.
- SimCLR model performance was matched by a fully supervised YOLOv11 classifier, pointing to the exceptional rate of progress in supervised learning



University of Cape Town
Department of Computer Science
TEL: +27 (0)21 650 2663
Website: https://sit.uct.ac.za/

Team
Catalina Thomson <altcat001@myuct.ac.za>`
Joseph Goldblatt <gldjos010@myuct.ac.za>
Max Elkington <elkmax001@myuct.ac.za>

Supervisors
Patrick Marais
<patrick@cs.uct.ac.za>
Jonathan Shock
<jonathan.shock@uct.ac.za>