UNDERWATER OBJECT DETECTION

Transformer vs. CNN: Benchmarking RT-DETR and YOLO Models on the UATD Dataset

Sercan KÜLCÜ

Department of Computer Engineering, Faculty of Engineering, Giresun University, Türkiye

17th International Academic Studies Congress (UBCAK) | October 23-24, 2025

Introduction & Motivation

- Forward-looking sonar (FLS) imagery suffers from low resolution, high noise, blurry boundaries, and sparse targets.
- Signal attenuation and seabed reverberation limit traditional pixel-based computer vision methods.
- Critical for marine resource exploration, subsea construction, and ecological monitoring.
- This study benchmarks Transformer-based RT-DETR-L against CNN-based YOLOv12n and YOLOv8n on the UATD dataset.

KEY DRIVERS

9,200 raw sonar (MFLS) images in UATD

10 distinct object categories

3 models compared: RT-DETR-L, YOLOv12n, YOLOv8n

Related Work

1 FLSD-Net / Fast-YOLOv5

Lightweight YOLOv8/v5 variants (Yang et al.) improve mAP and reach up to 200 FPS via noise-augmented training.

2 SonarNet

CNN-Transformer hybrid (He et al.) reaches 54.5% MIoU at 215.5 FPS for FLS segmentation.

3 FA-YOLO / DA-YOLOv7

Attention- and transfer-learning-enhanced YOLO variants reach up to 98–99% mAP on UATD/SCTD.

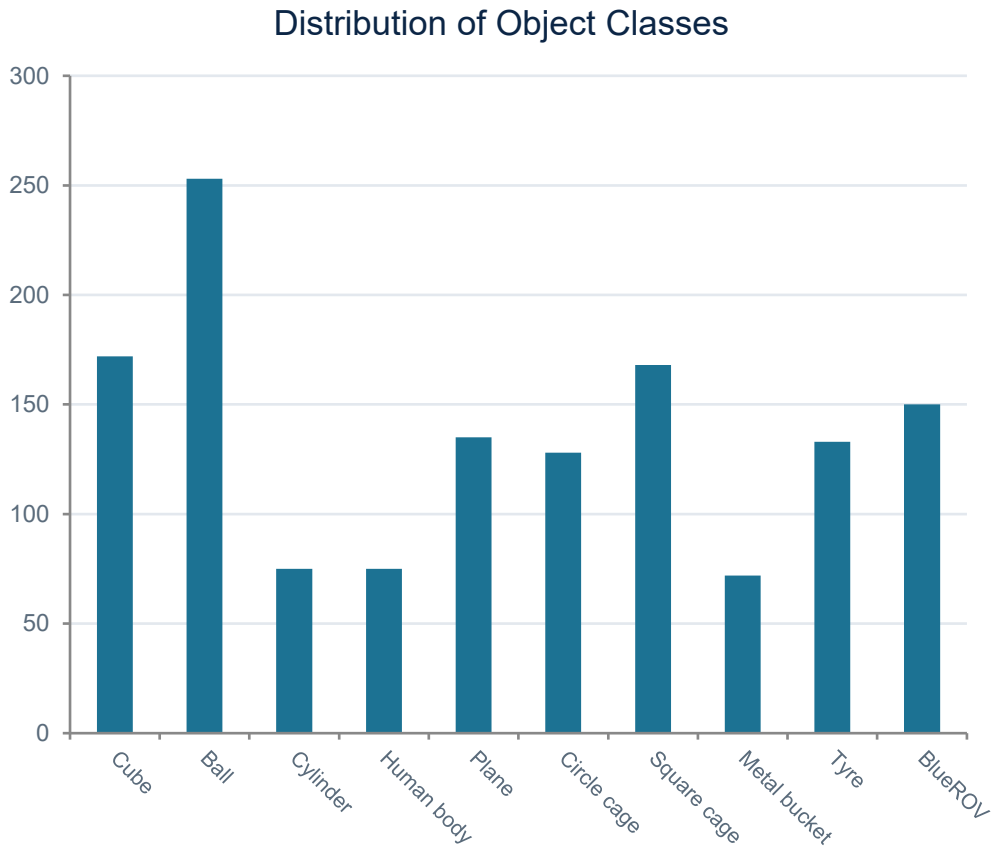
4 LS-DETR / YOLO-DCN

Lightweight Transformer and deformable-convolution detectors push accuracy gains of 2.8–10.2%.

Trade-off: Transformer models capture global context for high accuracy; CNN-based YOLO models prioritize computational efficiency for real-time use.

Dataset: UATD

- 9,200 raw sonar images from a Tritech Gemini 1200k multibeam forward-looking sonar (MFLS).
- Collected in lakes and shallow waters: Haoxin Lake, Maoming and Golden Pebble Beach, Dalian (China).
- BMP format, 1024×1428 resolution, 720/1200 kHz frequency, 5–25 m range.
- 7,600 training images, 800 test_1, 800 test_2; XML bounding-box labels.
- 10 object classes: cube, ball, cylinder, human body, tyre, circular cage, square cage, metal bucket, plane, blueROV.



95 images moved from training to test set to balance human-body and metal-bucket instance counts.

RT-DETR-L Training Setup

Framework	Ultralytics 8.3.167, pretrained COCO weights (926/941 layers transferred)
Epochs	20 (no early stopping, full training)
Image Size	640 × 640
Batch Size	4 (GPU memory limitation)
Optimizer	AdamW, lr = 0.000714, momentum = 0.9
Loss Functions	GIoU + Classification + L1
Augmentation	Mosaic = 1.0, HSV (h=0.015, s=0.7, v=0.4), FlipLR = 0.5
Hardware	NVIDIA GeForce RTX 3050 GPU (CUDA 11.8)

Trained model: 32,004,290 parameters (103.5 GFLOPs).

Results: Overall Model Comparison

Model	Params	GFLOPs	P	R	mAP50	mAP50-95	Inference (ms)
RT-DETR-L	**32.0M**	**103.5**	**0.953**	**0.964**	**0.963**	**0.535**	**24.1**
YOLOv12n	2.56M	6.3	0.906	0.906	0.953	0.481	3.5
YOLOv8n	3.01M	8.1	0.946	0.918	0.959	0.507	2.7

Table 2 — Performance comparison on the validation set (895 images, 1,362 instances).

RT-DETR-L

Highest accuracy

mAP50-95 = 0.535
Global context modeling

YOLOv8n

Best trade-off

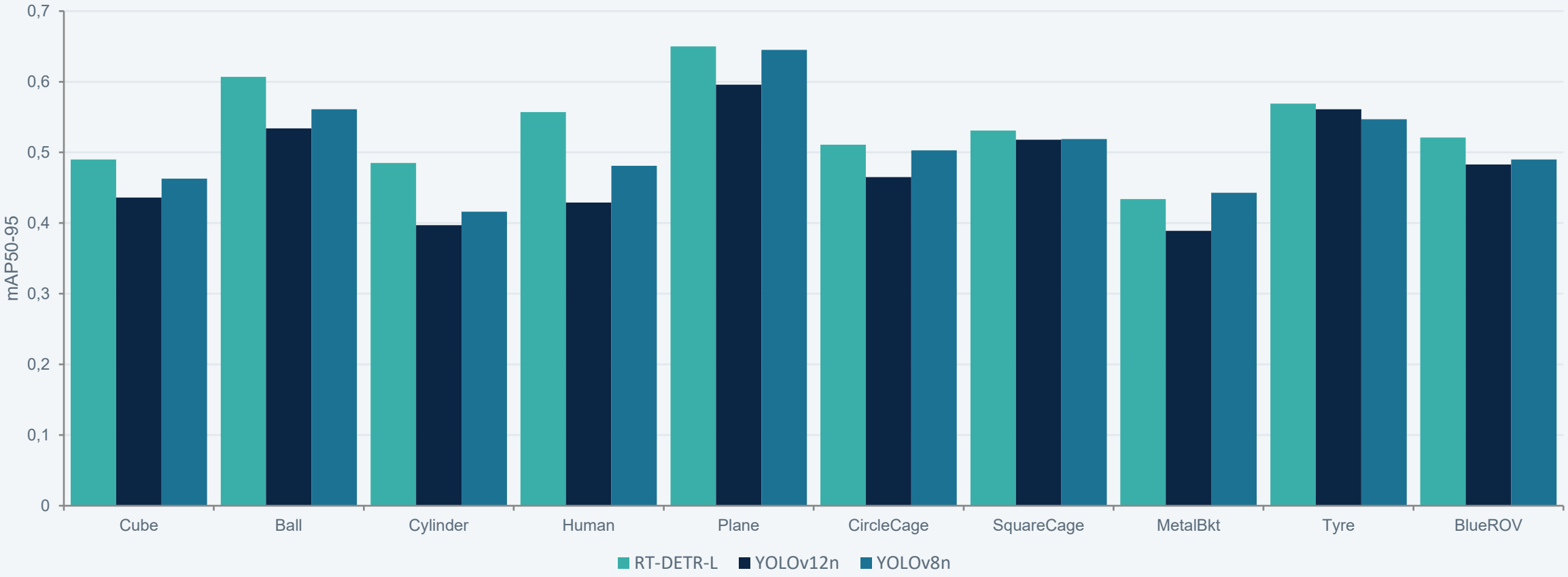
mAP50-95 = 0.507
2.7 ms inference

YOLOv12n

Fastest / lightest

6.3 GFLOPs
2.56M parameters

Class-Wise mAP50-95 Comparison



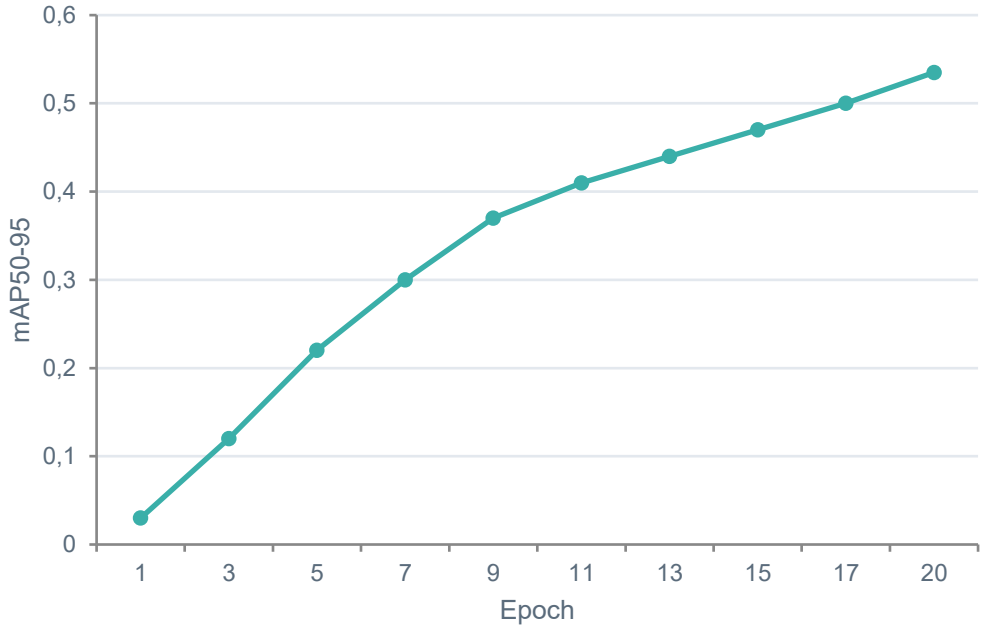
RT-DETR-L leads on plane (0.650) and ball (0.607) — classes with strong global acoustic signatures. YOLOv8n stays competitive with balanced per-class accuracy.

Confusion Matrix & Training Dynamics

Normalized Confusion Matrix — RT-DETR-L (diagonal)

Minor confusions: 19% of tyres → blueROV; 16% of square cages → circle cages (similar geometric profiles).

Training Progress — mAP50-95 (20 Epochs)



Rapid gains up to epoch 10; metrics plateau by epoch 20, indicating stable convergence.

Discussion

1 Accuracy vs. Speed Trade-off

RT-DETR-L's global context modeling drives the highest mAP50-95 (0.535), but at 103.5 GFLOPs and 24.1 ms inference — far heavier than YOLO variants.

2 Why RT-DETR-L Excels on Plane & Ball

Distinct structural outlines, acoustic shadows, and spherical reflections are captured by long-range dependency modeling in Transformers.

3 YOLOv8n's Balanced Profile

Best trade-off for resource-constrained underwater robotics — mAP50-95 of 0.507 with only 2.7 ms inference.

4 YOLOv12n's Efficiency Focus

Lightest and fastest (6.3 GFLOPs), but trails in precision on irregular shapes (cylinder 0.397, human body 0.429).

Conclusion

- RT-DETR-L achieves state-of-the-art accuracy (mAP50-95 = 0.535) for underwater sonar object detection via global context modeling.
- YOLOv8n and YOLOv12n offer far faster inference (2.7–3.5 ms) with lower computational cost, suited to resource-constrained AUVs.
- Model choice depends on application needs: precision-critical tasks favor Transformers; real-time onboard tasks favor YOLO.
- Future work: hybrid architectures combining Transformer accuracy with YOLO's lightweight efficiency.

0.535

Best mAP50-95
(RT-DETR-L)

2.7 ms
Fastest inference (YOLOv8n)

9,200
sonar images (UATD)

Contact: sercan.kulcu@giresun.edu.tr | ORCID: 0000-0002-4871-709X