ADVERSARIAL ATTACK ON MARITIME ENVIRONMENT

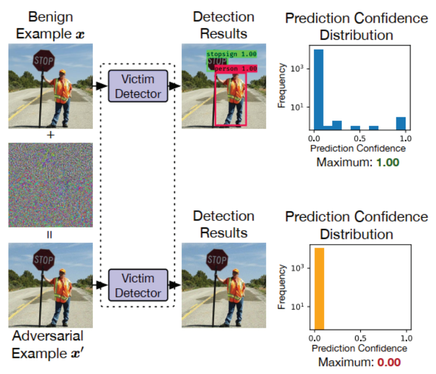
ZEYNEP YARADANAKUL | MD SALEH IBTASHAM | PHORNPHAWIT MANASUT

Problem Statements

Object detectors can be fooled by making small imperceptible changes to images.

These pose a challenge to various domain like autonomous cars and ships.

Adversarial attacks can cripple performance of the object detectors.



TEAM 7

Our Objectives

- Implement black box adversarial attack against object detection models
  - Train different YOLO versions on different combinations of maritime datasets
    - Implement TOG (Targeted Objectness Gradiant) adversarial attack
- Study the performance of transferability of Adversarial Attack between various maritime datasets and YOLO versions.

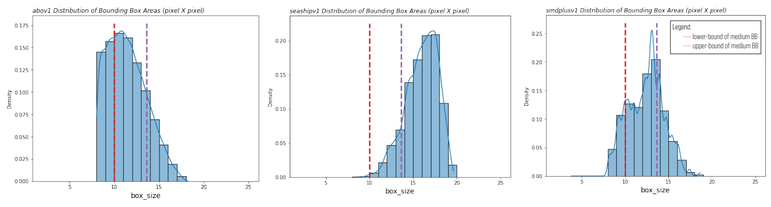
Datasets

Dataset Bounding Box (BB) Specifications:

Source Dataset	Mean BB Size (log_2)	No. Small BB	No. Medium BB	No. Large BB	No. Total BB
ABO	11.41	10047	17467	5713	33227
SS	15.98	15	1034	8172	9221
SMD	12.39	26921	97297	48369	172587

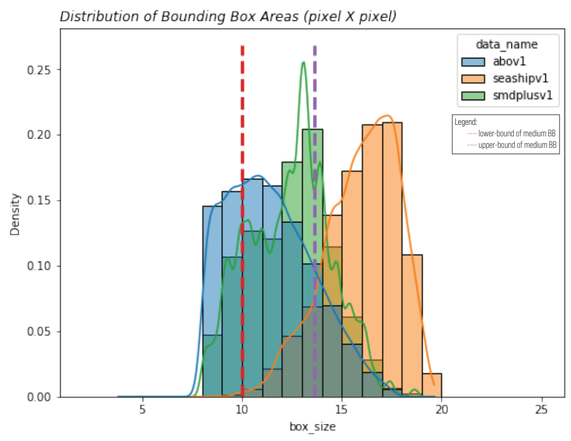
Datasets are abbreviated as follow: ÅboShips=ABO, SeaShips=SS, SMDPlus=SM

Dataset Bounding Box (BB) Distributions:



Object Size Visualization

Dataset Bounding Box (BB) Comparison:



Model & Dataset Compositions

Dataset ID	Source Dataset(s)
D1	ABO
D2	SS
D3	SMD
D4	ABO, SS
D5	ABO, SMD
D6	SS, SMD
D7	ABO, SS, SMD

- Model Training Scheme
  - 7 Dataset Combination (with 3 Datasets)
  - 3 YOLO Models (YOLOv3, YOLOv5, YOLOv8)
  - 21 possible combinations of models (3 (YOLO) X 7(Dataset) = 21)

Model ID	Source Architecture	Source Dataset	mAP50	mAP50-95
M1	YoloV3	D1	0.8154	0.4453
M2	YoloV3	D2	0.9926	0.8469
M3	YoloV3	D3	0.9900	0.9169
M4	YoloV3	D4	0.8738	0.5573
M5	YoloV3	D5	0.9722	0.8343
M6	YoloV3	D6	0.9899	0.8956
M7	YoloV3	D7	0.9753	0.8252
M8	YoloV5	D1	0.8190	0.4522
M9	YoloV5	D2	0.9930	0.8306
M10	YoloV5	D3	0.9895	0.9144
M11	YoloV5	D4	0.8715	0.5561
M12	YoloV5	D5	0.972	0.8396
M13	YoloV5	D6	0.9897	0.9006
M14	YoloV5	D7	0.9732	0.8294
M15	YoloV8	D1	0.8158	0.4518
M16	YoloV8	D2	0.9940	0.8454
M17	YoloV8	D3	0.9882	0.9229
M18	YoloV8	D4	0.8753	0.5596
M19	YoloV8	D5	0.9731	0.8436
M20	YoloV8	D6	0.9880	0.9033
M21	YoloV8	D7	0.9739	0.8338

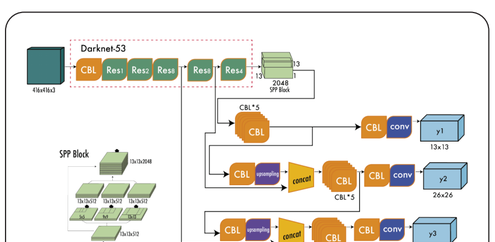
Model Architectures

YOLO architecture Properties:

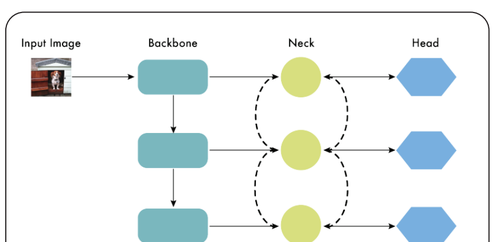
YOLO Version	Variant	Number of Layers	Parameters	Weight File Size (MB)
YOLOv3	Original	310	103754144	208
YOLOv5	Extra Large	493	97276448	195
YOLOv8	Extra Large	365	68229648	137

YOLOv3, YOLOv5, YOLOv8 selected because of extensive support of the Ultralytics library

YOLO architecture Designs:



Old Design (YOLOv3)



Modern Design (YOLOv5, YOLOv8)

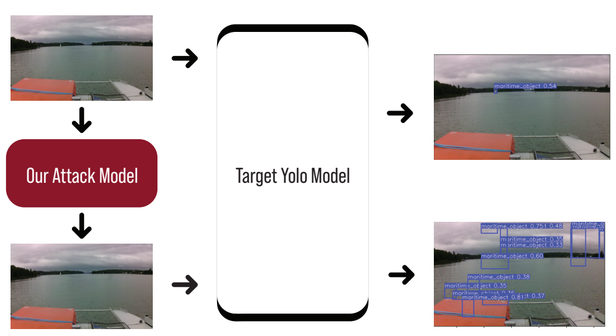
General Attack Flow

Results (Quantitative View)

mAP50 Reduction; target: predict, surrogate: perturb, gray boxes = white box attack

Target Model → / Surrogate Model ↓	M1	M2	M4	M8	M9	M11	M15	M16	M18
M1	0.725	0.004	0.347	0.278	0.002	0.201	0.284	0.001	0.207
M2	0.127	0.486	0.138	0.081	0.030	0.166	0.077	0.049	0.141
M4	0.476	0.041	0.656	0.289	0.009	0.245	0.321	0.014	0.240
M8	0.356	0.005	0.273	0.663	0.004	0.370	0.467	0.001	0.279
M9	0.097	0.071	0.149	0.087	0.326	0.260	0.083	0.096	0.180
M11	0.314	0.021	0.286	0.468	0.033	0.557	0.425	0.019	0.325
M15	0.341	0.005	0.238	0.371	0.002	0.275	0.705	0.002	0.310
M16	0.106	0.067	0.166	0.096	0.074	0.231	0.117	0.368	0.241
M18	0.344	0.032	0.297	0.383	0.020	0.343	0.512	0.034	0.583

Results (Qualitative View)



Targeted Objectness Gradient (TOG) Attacks

- A family of attacks that exploit various components of Object Detector Loss Function
- Attack selected for our experiements: TOG-Fabrication.
- Fabrication exploits entire loss function by multiplying it by -1 to compute the gradient.
- All TOG attacks conforms to the following general form:

$$x'_{t+1} = \prod_{x,\epsilon}[x'_t - \alpha\Gamma(\frac{\nabla L^*(x'_t; O^*, W)}{\nabla x'_t})]$$

Discussion

- As Expected, the higher the intersection between trained datasets the better the attack performs
- Models trained on dataset with smaller bounding box distribution are more susceptible to adversarial attacks.
- YOLOv5 performs the best against adversarial attacks, which may mean the higher number of layers in a model result in better robustness against adversarial attacks.

Future Work

- Validate surrogate model performance against other datasets
- Multi Label cross-data transferability attack performance
- Meta-surrogate training and attack performance on multi & Single labels