# Self-driving Vessels: YOLOv5 Approach for Water Surface Object Detection

**T. R. D. Sá**[1], **C. M. S. Figueiredo**[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA-EST)
CEP 69050-020 – Manaus – AM – Brasil

trs.eng17@uea.edu.br, cfigueiredo@uea.edu.br

***Abstract.*** *The use of Computer Vision techniques in Water Surface Object Detection has risen as a strong trend in autonomous vessels context. We propose an YOLOv5 algorithm performance avaliation for water surfaces objects. Following, we compare it with the benchmark results of others 17 classical detectors. This work uses an annotated image dataset available from a benchmark dataset, named WSSOD, characterized by being public, wide (7,467 images, 14 categories and differents capture conditions) and specialized in water surface objects. YOLOv5 reached a mAP of 76.3 %, outstanding in 11.3 % the mAP obtained by CRB-Net detector on the WSODD benchmark dataset.*

***Resumo.*** *A aplicação de técnicas de Visão Computacional para detecção de objetos em superfície de água tem destacado-se como uma forte tendência no contexto de embarcações autônomas. Este trabalho apresenta uma avaliação de performance do algoritmo YOLOv5 para detecção de objetos localizados em superfície de água. Em seguida, o compara com a performance de outros 17 detectores clássicos. É utilizado um conjunto de imagens anotadas e disponibilizadas em uma base imagens de referência - WSODD, sendo essa caracterizada por ser pública, abrangente (7.467 imagens, 14 categorias e diferentes condições de captura) e especializada em objetos localizados em superfície de água. Finalmente, YOLOv5 obteve um mAP igual à 76.3 %, tendo superado em 11.3 % o mAP obtido pelo detector CRB-Net na mesma base de referência WSODD.*

## 1. INTRODUÇÃO

Em conformidade com as regulamentações da COLREGs[1] (*Convention on the International Regulations for Preventing Collisions at Sea, 1972*), evitar colisões com objetos estáticos ou dinâmicos, sem que haja intervenção humana, de maneira geral, é um dos requisitos mais importantes para a viabilização de embarcações autônomas [Gu et al. 2019]. Nesse cenário, elas podem ser desenvolvidas como uma infraestrutura flexível, isto é, uma ponte temporária ou uma estação flutuante, evitando obstáculos à medida que navega pela superfície da água de forma autônoma [Wang et al. 2020]. Além disso, podem também ser utilizadas em contextos onde há uma demanda por vigilância militar [Prasad et al. 2017], controle de tráfego marítimo [Prasad et al. 2020] e monitoramento ambiental [Bloisi et al. 2012, Bloisi et al. 2011]. Esses exemplos representam algumas

[1]Para mais detalhes, acessar: https://www.imo.org/en/OurWork/Safety/Pages/Preventing-Collisions.aspx

das muitas aplicações possíveis à medida que métodos de Visão Computacional e técnicas de Inteligência Artificial, com vistas à detecção de objetos localizados em superfície de água, são devidamente implementados [Xu et al. 2017].

Existem muitos algoritmos especializados em detecção de objetos, tais como *SSD* [Liu et al. 2015], *RetinaNet* [Lin et al. 2017], *Fast-RCNN* [Girshick 2015], *Mask-RCNN* [He et al. 2017], *YOLO* [Li et al. 2020] e entre outros. Especialmente em relação ao algoritmo YOLO, sua quinta versão, YOLOv5, encontra-se em estado da arte e tem se destacado pela sua praticidade e performance frente a outros detectores clássicos. Além disso, foi pouco explorado para contextos de objetos localizados em superfície de água. Por último, YOLOv5 incorpora técnicas de aumentação de imagens que são aplicadas em tempo real de treinamento, provendo modelos robustos e que generalizam melhor. Em virtude desses motivos, foi escolhido como detector de referência a ser avaliado neste trabalho.

Ainda que haja promissores detectores de objetos em estado da arte disponíveis conforme mencionado acima, há uma significativa escassez de abordagens baseadas em arquiteturas convolucionais que são especializadas na detecção de objetos localizados em superfície de água [Prasad et al. 2017]. Isso ocorre principalmente em virtude de existirem poucas bases de imagens disponíveis publicamente que retratam objetos em superfície de água [Zhou et al. 2021]. Das poucas bases disponíveis, a maioria é caracterizada por não ser representativa em relação à quantidade de instâncias por categoria. Além disso, muitas delas apresentam pouca variedade no que diz respeito às condições de captura das imagens. *ImageNet* [Deng et al. 2009], *MS COCO* [Lin et al. 2014] e *Places 2* [Zhou et al. 2018] são exemplos de algumas bases de imagens públicas que representam os problemas citados anteriormente, pois, apesar de possuírem grande volume de imagens, são bastantes escassas no que tange a objetos em superfície de água. Esse cenário representa um notável problema, pois impacta negativamente para o fomento ao desenvolvimento de técnicas de detecção de objetos em superfície de água.

Para testar, comparar e otimizar os métodos de detecção para o contexto de objetos localizados em superfície de água, é fundamental que volumosas bases de imagens anotadas estejam disponíveis publicamente. Caso essa condição não seja tangível, é possível optar pela abordagem da criação de uma base de imagens própria, sendo essa alternativa, a priori, muitas vezes inviável a curto ou médio prazo. De todo modo, tendo sido obtida uma base de referência, diferentes abordagens de detecção de objetos são testados, permitindo a comparação e otimização de performance em métricas de acurácia e tempo de inferência.

Recentemente, foi disponibilizada uma base de imagens anotadas que retratam, com notável qualidade e variedade, objetos localizados em superfície de água. Essa base, intitulada como *WSODD* (*Water Surface Object Detection Dataset*) [Zhou et al. 2021], mitiga os pontos levantados anteriormente e será utilizada como referência para comparar a performance do detector YOLOv5 com diferentes detectores clássicos de objetos. Considerando o que foi exposto acima, as principais contribuições deste artigo são:

(I) Avaliar a performance da versão mais recente do detector YOLO (YOLOv5) na base WSODD, considerando as principais métricas de avaliação utilizadas em modelos de detecção de objetos. Ao final, verificou-se que o modelo treinado para detectar objetos

em superfície de água alcançou ótima capacidade de generalização frente a diferentes categorias.

(II) Comparar a performance do detector YOLOv5 com outros diferentes métodos de detecção apresentados no trabalho de [Zhou et al. 2021], utilizando a mesma base de imagens WSODD. Ao final, observou-se que o método YOLOv5 superou significativamente todas as outras abordagens comparadas.

Sob a ótica da proposta deste trabalho, Seção 2 apresenta brevemente aspectos essenciais de fundamentação teórica que norteiam as bases deste trabalho. Em seguida, Seção 3 descreve sobre os trabalhos relacionados e como este trabalho destaca-se em relação aos demais. Seção 4 contempla os métodos empregados para alcançar os resultados obtidos. Seção 5 apresenta uma discussão sobre os resultados alcançados. Por fim, Seção 6 apresenta considerações finais sobre o que foi desenvolvido.

## 2. FUNDAMENTAÇÃO TEÓRICA

Essa seção apresenta, resumidamente, os principais conceitos teóricos que serão utilizados ao decorrer do trabalho proposto.

### 2.1. Redes Neurais Convolucionais

Redes Neurais Convolucionais (CNNs) [LeCun et al. 2001] são arquiteturas de Aprendizado Profundo comuns em algoritmos de Visão Computacional, especialmente em detectores de objetos. Em termos gerais, CNNs são redes neurais simples que utilizam, em pelo menos uma de suas camadas, uma operação conhecida por convolução [Goodfellow et al. 2016]. Além disso, CNNs possuem camadas que são imprescindíveis para o processamento de tarefas cujos dados de entrada são imagens, tais como, por exemplo, filtros detectores de características, mapas de características, e entre outras. Por meio dessas camadas, é possível implementar algoritmos robustos para tarefas preditivas inseridas no contexto de Visão Computacional, em especial à detecção de objetos.

### 2.2. Técnicas utilizadas em detecção de objetos

Detecção de objetos é um ramo da Visão Computacional que é responsável por inferir a localização de um objeto em uma dada imagem, junto com a sua respectiva classe [Zhao et al. 2018]. De maneira geral, os métodos mais comuns que auxiliam nessa tarefa podem ser categorizados em dois principais tipos: métodos de passada única e métodos de passada dupla.

Métodos de passa única priorizam o tempo de resposta da inferência. Os detectores YOLO [Li et al. 2020], SSD [Liu et al. 2015] e RetinaNet [Lin et al. 2017] são exemplos desse método. Por outro lado, métodos de passada dupla priorizam a acurácia da detecção. Alguns algoritmos destacam-se para esse método, tais como *Fast R-CNN* [Girshick 2015], *Faster R-CNN* [Ren et al. 2015], *Mask-RCNN* [He et al. 2017] e entre outros.

Como exemplo, tem-se o algoritmo YOLO (*You Only Look Once*), que é um método baseado em Redes Neurais Convolucionais que realiza detecção de objetos com uma passada única. Ele foi idealizado para sistemas que necessitam realizar inferências em um intervalo de tempo curto, enquanto mantém uma acurácia alta. Para isso, ele

aborda a detecção de objetos como um único problema de regressão, onde, diretamente a partir dos *pixels* de uma dada imagem, extrai as coordenadas das caixas delimitadoras, juntamente com as probabilidades das classes.

## 3. TRABALHOS RELACIONADOS

Detecção de objetos em superfície de água, especialmente em navegações marítimas, compreende um cenário de notável desafio em virtude da escassez mesclada com a complexidade das imagens observadas. A seguir são apresentados alguns trabalhos de destaque relacionados ao tema de detecção de objetos no contexto de embarcações autônomas.

A base de imagens mencionada na Seção 1 foi introduzida pelo trabalho de [Zhou et al. 2021] . Nesse trabalho, destaca-se a base de imagens WSODD como uma contribuição muito significativa para o cenário de detecção de objetos em superfície de água. Ela é caracterizada por possuir uma quantidade significativa de instâncias por categorias, sendo composta por 7.467 imagens, diversificadas em 3 ecossistemas distintos, isto é, oceanos, lagos e rios; 3 condições de iluminação distintas - ao amanhecer, ao entardecer e ao anoitecer; 3 condições climáticas, isto é, dias ensolarados, nublados e nebulosos. Ao todo, existem 14 classes e 21.911 objetos anotados. Além disso, destacam-se também os testes comparativos que foram realizados com diferentes métodos de detecção de objetos. Ao todo, 17 detectores foram testados e avaliados em função da métrica *mAP* (*Mean Average Precision*). Tanto a base de imagens quanto as métricas obtidas foram utilizadas como referência para comparar com os resultados obtidos neste presente trabalho.

No trabalho de [Prasad et al. 2017], são exploradas diferentes abordagens para a detecção de objetos localizados na superfície do mar, visando o cenário de embarcações autônomas. Em síntese, são apresentados detectores do tipo passada dupla que contemplam três etapas, a saber, (i) detecção de horizontes, (ii) subtração de segundo plano e (iii) segmentação de primeiro plano. *Singapore Maritime Dataset* [2] é a base de imagens utilizada para avaliar a performance dessas diferentes abordagens de detecção de objetos. Por fim, observa-se que essa base compreende apenas ambientes marítimos.

Por último, os autores em [Ribeiro et al. 2019] apresentam uma base de imagens anotadas, capturadas a partir de VANTS (Veículos Aéreos Não Tripulados). A principal finalidade é de possibilitar vigilância em um contexto marítimo. Para isso, o trabalho propõe alguns experimentos voltados à detecção de embarcações. Para isso, utiliza detectores clássicos do tipo passada dupla, em especial os métodos *R-CNN* e *Blob*, além de abordagens não supervisionadas. De maneira geral, nota-se que as imagens capturadas são restritas a ambientes marítimos, especializadas apenas para embarcações e em escalas muito pequenas.

Com exceção do primeiro trabalho mencionado, o restante compartilha uma característica em comum, isto é, as bases de imagens utilizadas não são representativas de objetos localizados em superfície de água. Em síntese, as imagens retratam, majoritariamente, apenas embarcações, não abrangendo, todavia, outros possíveis tipos de objetos, tais como rochas, vegetação, portos, entulhos, animais, plataformas de embarque, boias, pontes e entre outros. Além disso, as imagens retratam apenas embarcações em cenários

[2]Mais detalhes sobre essa base de imagens, acessar: `https://sites.google.com/site/dilipprasad/home/singapore-maritime-dataset`

marítimos, não contemplando outros ecossistemas aquáticos, isto é, lagos, lagoas, rios, geleiras e entre outros. Por fim, este trabalho destaca-se dos demais por avaliar e comparar, em primeira mão, a performance de um detector em estado de arte, YOLOv5, em uma base de imagens com notável representatividade em diferentes objetos e diferentes ecossistemas aquáticos.

## 4. MATERIAIS E MÉTODOS

Com vistas ao objetivo de avaliar e comparar a performance do algoritmo YOLOv5 para detecção de objetos localizados em superfície de água, foram considerados três principais etapas para composição do experimento, isto é, (i) adequação das anotações da base de imagens, (ii) definição do formato de validação do modelo e (iii) definição do ambiente de desenvolvimento. A seguir, são descritas em detalhes cada uma dessas etapas.

A base de imagens WSODD foi criada com o propósito de contribuir para o contexto de detecção de objetos localizados em superfície de água. É uma base com imagens anotadas, fornecendo informações tais como coordenadas da localização objeto, categoria e resolução. Todas essas anotações são dados imprescindíveis para o treinamento de modelos orientados à detecção de objetos. Ainda, é composto por uma considerável quantidade de imagens distribuídas entre diferentes ecossistemas aquáticos, estações do dia e condições climáticas. Enfim, até o presente momento, essa é a base disponibilizada publicamente que possui a maior abrangência em função das características acima mencionadas.

As anotações das imagens pertencentes à base WSODD foram disponibilizadas em formato PASCAL VOC [Everingham et al. 2015], não sendo possível, portanto, utilizá-las para o treinamento de modelos a partir do método YOLO, pois este requer que as anotações estejam em formato YOLO. Dessa forma, foi realizado a implementação de um algoritmo, escrito em Python, apto a converter as anotações do formato de origem, PASCAL VOC, para o formato de destino, YOLO.

No que tange a validação do modelo, foi utilizado o método *holdout* para a validação cruzada do experimento. Dessa forma, foi mantido a mesma proporção utilizada em [Zhou et al. 2021], ou seja, 70 % das imagens foram destinadas para o conjunto de treinamento, enquanto que o restante, 30 %, foram utilizadas no conjunto de teste. Ressalta-se que a seleção delas, em ambos os conjuntos, foi realizada de forma aleatória, mantendo uma distribuição representativa das classes em cada conjunto.

Em relação aos detalhes de implementação, para realizar o experimento com o método YOLOv5, foi utilizado uma estação de trabalho (computador) com as seguintes configurações: I. Sistema operacional Ubuntu 20.04; II. Processador Intel Core i7 de 7ª Geração; III. Memória RAM 16GB; IV. Armazenamento HD de 1TB; V. Placa de vídeo dedicada NVIDIA GeForce GTX 1060 com 6GB.

As etapas de treinamento e validação foram implementadas em contêineres [3] . Isso viabilizou o encurtamento no tempo necessário para desenvolvimento do modelo, pois os contêineres disponibilizados com o algoritmo YOLOv5 caracterizam-se por possuírem

[3]Referências sobre o uso do YOLOv5 com Docker, consultar: `https://docs.ultralytics.com/environments/Docker-Quickstart/`

todas as dependências de bibliotecas empacotadas em uma única imagem do sistema operacional.

Por fim, para os hiper-parâmetros utilizados arquitetura convolucional do detector YOLOv5, foi definido uma taxa de aprendizado de 0.01, momento igual à 0.937 e deicamento de peso de 0.0005. Em seguida, definiu-se o tamanho de lote igual à 16, com número épocas igual à 300. Além disso, para a otimização da função de perda, foi escolhido o otimizador SGD [4] (*Stochastict Gradient Descent*). Observa-se que, como etapa de preprocessamento, as imagens de entrada foram redimensionadas para 416 x 416 pixels. As escolhas desses hiper-parâmetros foram baseadas por meio da observação empírica, isto é, tentativa e erro, além de também se basear nos valores padrões utilizados pelo método YOLOv4 na base de imagens PASCAL VOC.

## 5. RESULTADOS

Essa seção disponibiliza os resultados e discussões que dizem respeito ao modelo treinado a partir do algoritmo YOLOv5, com as imagens e anotações oriundas da base de referência WSODD. A seção é divida em duas partes: (i) discussões em torno das métricas de avaliação de performance do modelo e (ii) comparações entre a performance obtida nesse modelo com a performance de 17 outros detectores apresentados em [Zhou et al. 2021].

### 5.1. Avaliação de performance do modelo

A curva da função de perda fornece um "retrato"do processo de treinamento do modelo e a direção em que a rede neural está aprendendo. Figura 1-a, Figura 1-b e Figura 1-c ilustram essa curva para três diferentes variáveis, isto é, *box_loss*, *obj_loss* e *cls_loss*. Observa-se que, à medida que o número de épocas aumenta, há uma convergência relativamente rápida no sentido da redução do erro. Por fim, a Figura 1-f, Figura 1-g e Figura 1-i ilustram o decaimento do erro no conjunto de teste para as mesmas três variáveis mencionadas acima.

A precisão e revocação são métricas de avaliação de performance comuns em modelos de detecção de objetos, pois conseguem medir a qualidade das classificações em termos de verdadeiros positivos (TP), falsos positivos (FP) e falsos negativos (FN). Figura 1-d e Figura 1-e ilustram como as métricas de precisão e revocação evoluíram ao longo do treinamento.

Ainda, observa-se que o modelo conseguiu aprender a detectar, com notável performance, objetos de diferentes categorias, pois houveram incrementos positivos na precisão e revocação à medida em que os falsos positivos e falsos negativos foram reduzidos.

A Figura 2 sintetiza o resultado das classificações para cada categoria por meio de uma matriz de confusão. Por meio dela, observa-se que as categorias que tiveram a maior porcentagem de instâncias classificadas erroneamente foram as classes *grass* (grama) e *mast* (mastro). A razão disso pode ser em virtude da baixa quantidade de imagens e instância que representam essas classes no conjunto de treino.

Por fim, a Figura 3 ilustra imagens pertencentes ao conjunto de teste que foram submetidas ao processo de inferência para detecção de objetos localizados em superfície de água.

[4]Mais informações em: `https://en.wikipedia.org/wiki/Stochastic_gradient_descent`

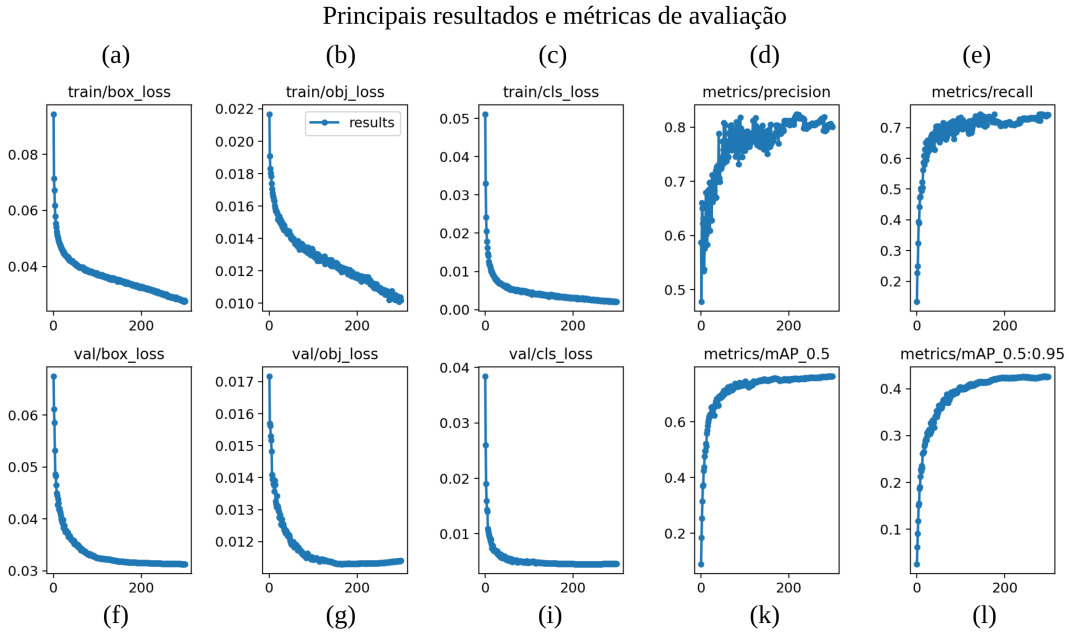


**Figura 1. Visão geral dos resultados do treinamento do modelo baseado no detector YOLOv5. Os gráficos (a), (b), (c), (f), (g) e (i) ilustram a curva da função de perda para 3 diferentes variáveis, considerando tanto o conjunto de treino (*train*), quanto o conjunto de teste (*val*). Os comportamentos da precisão (*precision*) e revocação (*recall*) são ilustrados pelos gráficos (d) e (e). Por fim, as curvas ilustradas nos gráficos (k) e (l) fornecem resultados de mAP do modelo.**

## 5.2. Comparação entre modelos de referência

Ao comparar métodos tradicionais de detecção de objetos com métodos baseados em arquiteturas convolucionais profundas, verifica-se que estes últimos têm alcançando uma maturidade significativa em um curto espaço de tempo, pois eles possuem capacidade de aprendizado muito superior aos métodos clássicos. Isso é evidenciado pelos resultados obtidos e disponibilizados na Tabela 1, onde são avaliados diversos modelos de detecção de objetos em função da métrica mAP. Todos os modelos apresentados nessa tabela foram treinados a partir da base de imagens WSODD, utilizando o mesmo critério de validação mencionado na seção 4.2. A seguir, são discutidos os resultados alcançados comparando-os com os resultados obtidos a partir do algoritmo YOLOv5.

Os dois primeiros modelos (DPM e RANSAC-SVM) correspondem métodos de detecção de objetos baseados em técnicas clássicas de aprendizado de máquina. Observa-se um mAP baixo comparados os outros modelos. Todos os outros modelos restantes apresentados foram implementados baseadas em técnicas de aprendizado profundo. Em especial, os métodos Yolov3-2SMA e ShipYolo foram desenvolvidos para o cenário de detecção de objetos localizados em superfície de água. Em destaque, foi adicionado a contribuição deste trabalho em que o modelo treinado a partir do método YOLOv5 obteve um mAP de 76.3 %, considerando a base de imagens WSODD. Desse modo, a partir dos resultados obtidos, verificou-se que esse modelo alcançou notável capacidade de generalização para objetos localizados em superfície de água.

O ganho de performance do método YOLOv5 em comparação aos outros modelos

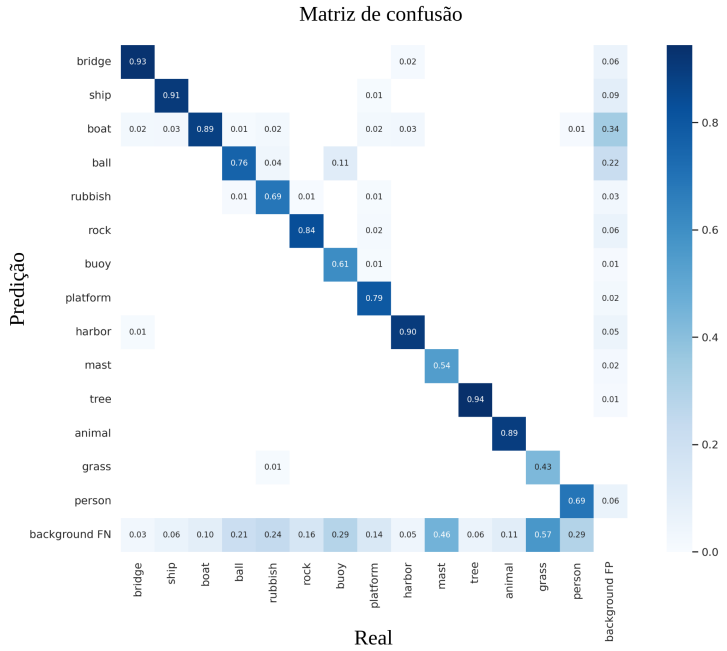


**Figura 2. Matriz de confusão das classificações obtidas pelo modelo.**

**Tabela 1. Comparação de desempenho entre 18 distintos métodos de detecção de objetos.**

| Método | mAP (%) |
|---|---|
| DPM | 21.9 |
| RANSAC-SVM | 27.1 |
| Faster R-CNN | 32.3 |
| Mask R-CNN | 35.7 |
| Cascade R-CNN | 41.1 |
| TridentNet | 62.2 |
| EfficientDet | 31.3 |
| RetinaNet | 27.9 |
| RFBNet | 35.7 |
| M2Det | 39.3 |
| SSD | 41.5 |
| CenterNet | 53.5 |
| Yolov3 | 56.1 |
| Yolov3-2SMA | 56.9 |
| Yolov4 | 57.2 |
| ShipYolo | 58.4 |
| CRB-Net | 65.0 |
| **Yolov5** | **76.3** |

apresentados pode ser justificado pela utilização de técnicas de aumentação de imagens que são aplicadas em lote durante a etapa de treinamento, corroborando para o desenvolvimento de um modelo que generaliza melhor frente a diferentes condições de capturas

das imagens disponíveis no dataset WSODD.

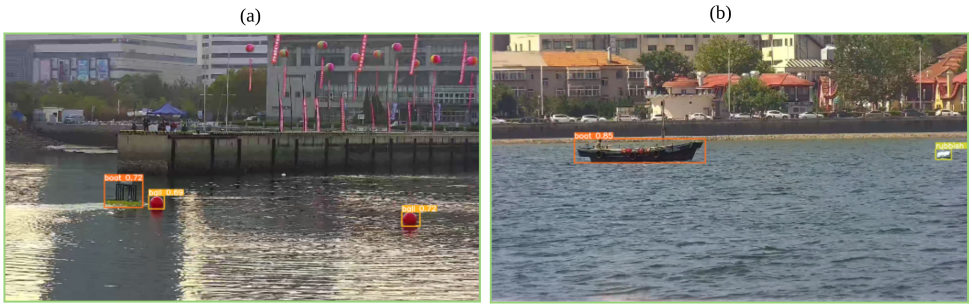


**Figura 3. Exemplos de objetos detectados após o treinamento do modelo. Na figura (a) nota-se a presença de caixas delimitadoras em volta de um barco e de uma bola (ball). Na figura (b), são detectados dois objetos: um barco (boat) e um entulho (rubbish).**

## 6. CONCLUSÃO

Neste trabalho, foi apresentado uma avaliação de desempenho do algoritmo YOLOv5 para detecção de objetos localizados em superfície de água, em diferentes condições de captura. Em seguida, sua performance foi comparada com outros 17 métodos clássicos de detecção de objeto. Tanto na avaliação quanto na comparação, foi utilizada a base de imagens de referência WSODD, caracterizada por ser pública, abrangente e especializada em imagens voltadas para objetos em superfície de água. A partir dos resultados obtidos, verificou-se que o modelo treinado alcançou notável capacidade de generalização frente a diferentes categorias de objetos. Por fim, nessa linha de pesquisa, trabalhos futuros podem ser realizados tomando como ponto de partida os estudos comparativos realizado com o detector YOLOv5.

## Referências

Bloisi, D., Iocchi, L., Fiorini, M., and Graziano, G. (2012). Camera based target recognition for maritime awareness. In *2012 15th International Conference on Information Fusion*, pages 1982–1987.

Bloisi, D. D., Iocchi, L., Fiorini, M., and Graziano, G. (2011). Automatic maritime surveillance with visual target detection.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Girshick, R. B. (2015). Fast R-CNN. *CoRR*, abs/1504.08083.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gu, Y., Góez, J., Guajardo, M., and Wallace, S. (2019). Autonomous vessels: State of the art and potential opportunities in logistics. *SSRN Electronic Journal*, (2019/6).

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2017). Mask R-CNN. *CoRR*, abs/1703.06870.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (2001). Gradient-based learning applied to document recognition. In *Intelligent Signal Processing*, pages 306–351. IEEE Press.

Li, X., Tian, M., Kong, S., Wu, L., and Yu, J. (2020). A modified yolov3 detection method for vision-based water surface garbage capture robot. *International Journal of Advanced Robotic Systems*, 17(3):1729881420932715.

Lin, T., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017). Focal loss for dense object detection. *CoRR*, abs/1708.02002.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., and Berg, A. C. (2015). SSD: single shot multibox detector. *CoRR*, abs/1512.02325.

Prasad, D. K., Dong, H., Rajan, D., and Quek, C. (2020). Are object detection assessment criteria ready for maritime computer vision? *IEEE Transactions on Intelligent Transportation Systems*, 21(12):5295–5304.

Prasad, D. K., Rajan, D., Rachmawati, L., Rajabally, E., and Quek, C. (2017). Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):1993–2016.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497.

Ribeiro, R., Cruz, G., Matos, J., and Bernardino, A. (2019). A data set for airborne maritime surveillance environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2720–2732.

Wang, W., Shan, T., Leoni, P., Fernández-Gutiérrez, D., Meyers, D., Ratti, C., and Rus, D. (2020). Roboat II: A novel autonomous surface vessel for urban environments. *CoRR*, abs/2007.10220.

Xu, Q., Yang, Y., Zhang, C., and Zhang, l. (2017). Deep convolutional neural network-based autonomous marine vehicle maneuver. *International Journal of Fuzzy Systems*, 20.

Zhao, Z., Zheng, P., Xu, S., and Wu, X. (2018). Object detection with deep learning: A review. *CoRR*, abs/1807.05511.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

Zhou, Z., Sun, J., Yu, J., Liu, K., Duan, J., Chen, L., and Chen, C. L. P. (2021). An image-based benchmark dataset and a novel object detector for water surface object detection. *Frontiers in Neurorobotics*, 15.