YOLOv5-FRNet: A Real-time Drowning Detection Method with Multi-Model Cascade

Yitong Zhou[1], Jingjing Wang[1], Lu Li[1], Wanwan Wang[2]

[1] School of Artificial Intelligence and Big Data, Henan University of Technology, Henan, Zhengzhou, China

[2]IFLYTEK CO.LTD, Anhui, Hefei, China

Abstract: Anti-drowning detection is of great significance to ensuring the safety of public waters, but the existing vision-based detection methods have problems such as high target miss detection rate and insufficient extraction of behavioral features in complex scenarios, and there is a lack of public anti-drowning detection datasets. To this end, this experiment collected and disclosed 8518 drowning prevention detection dataset and use LabelImg for image annotation. Then, a real-time drowning detection method with multi-model cascade was designed and implemented. First, the effects of Faster-RCNN and YOLOv5 methods in drowning detection were compared, and the results were cascaded to form YOLOv5-FRNet. The experimental results show that in terms of recall, YOLOv5-FRNet is 0.603, which is 0.01 and 0.113 higher than Faster-RCNN and YOLOv5, respectively. In terms of comprehensive detection performance, the IoU=0.5 index: YOLOv5-FRNet (0.774) is optimized with 1.3% and 27.7% compared with Faster-RCNN (0.764) and YOLOv5 (0.606) respectively; the IoU=0.5:0.95 index: YOLOv5-FRNet model (0.603) is 25.6% compared with Faster-RCNN (0.480), significantly better than YOLOv5 (0.388). The methods proposed in this article can be applied to swimming pools and other places, providing a reference for high-precision anti-drowning detection.

Keywords: Target Detection; Anti-Drowning Detection; YOLOv5; Faster-RCNN

1. Introduction

With the popularity of water recreation activities around the world, drowning accidents have become an important issue that threatens public safety. According to data from the World Health Organization (WHO), drowning is the third most common reason for unintentional fatalities [1]. Drowning has become one of the world's largest preventable, neglected and urgent public health problems. In order to attract attention from the international community, the World Health Organization has successively issued two official documents, "Guidelines for the Prevention of Drowning" and "Global Drowning Report", and has been updated year by year according to the situation [2]. Traditional manual monitoring methods are limited by labor costs and attention fatigue, making it difficult to achieve accurate warnings all day. In this context, intelligent anti-drowning system based on computer vision has gradually become a research hotspot. Its core lies in real-time identification of abnormal behaviors of drowning people (such as limb stiffness, head sinking, etc.) through target detection technology to win prime time for rescue. The existing drowning detection methods fall into two main categories. The first is a wearable sensor-based method and the second is a vision-based method [3]. Sensor-based methods use wearable devices (such as smart bracelets) to monitor physiological signals, but there are problems such as strong invasiveness and easy to fall off; based on visual analysis methods, early research mostly uses background subtraction, optical flow method or OpenPose attitude estimation, but it is lack of robustness to complex water environments (waves, light changes, multi-person occlusion).

In the design and optimization of drowning detection system, the core goal focuses on building a robust detection framework with high recall and high Average Precision (AP) to meet the needs of life safety monitoring in complex water environments. High recall rates are designed to minimize the risk of missed detection and ensure sensitive capture of limb movements of drowning people, while high average accuracy requires the system to maintain a low false alarm rate under complex interference conditions and reduce the ineffective scheduling of emergency response resources. In order to ensure high recall and high precision detection, this paper proposes a detection method that cascades the YOLOv5 and Faster-RCNN models to ensure the safety of people in the waters.

2. Current Status of Domestic and Foreign Research

In response to the technical problems such as small target size and susceptibility to wave occlusion in water human target detection, domestic and foreign scholars have proposed multi-level solutions. Wang [4] develop a reservoir drowning warning system based on the YOLOv4 framework, analyzes the dynamic spatial relationship between swimmers and dangerous areas through space-time context modeling, and combines long and short-term memory network (LSTM) to model behavior sequences, achieving a detection accuracy of 91.33% and a recall rate of 92.88%. In the field of posture behavior analysis, Jing et al. [5] proposed the SAG-Mask improvement architecture, constructed a spatial attention-guided segmentation branch based on Mask R-CNN, and increased the human contour segmentation accuracy by 15-20% by integrating timing inter-frame motion characteristics. Peng et al. [6] uses a joint-level feature pyramid network to optimize the multi-scale feature fusion mechanism, so that the drowning detection rate in complex swimming pool scenarios reaches 93.3%. It is worth noting that Yu and Yang [7] broke through the traditional judgment method based on the head position rules, innovatively used OpenPose to extract 17-node human key points to build a space-time distance matrix, and realized behavioral similarity measurement through graph convolution network, increasing the accuracy of drowning judgment to 95%. In the direction of multimodal perception technology, Chen et al. [8] constructed a fusion detection system based on SSD-OpenCV, and used a monocular visual distance measurement algorithm to calculate the Euclidean distance between the personnel and the dangerous area in real time to achieve active early warning response. Xie et al. [9] integrates Zigbee-iBeacon hybrid positioning technology, and establishes a swimming pool health risk assessment model through the fusion of physiological parameters (heart rate, blood oxygen saturation) and three-dimensional positioning data. In terms of detection model optimization, Liu et al. [10] innovatively constructed the YOLOv5-Dy-GBCA model. By introducing the GhostBottleneck lightweight module and coordinate attention mechanism to reconstruct the Backbone network, effectively enhancing the multi-level feature representation ability of small-scale targets (such as the head area of the drowning person), and combining dynamic detection heads to achieve adaptive optimization of spatial positioning accuracy, achieving 91.5% mAP on the self-built drowning data set, an increase of 8.2 percentage points from the benchmark model.

Although existing research has made some progress in the field of drowning prevention detection, most methods are still limited by the inherent shortcomings of a single model: high-precision models (such as Faster-RCNN) are difficult to meet real-time requirements, while lightweight models (such as YOLOv5) have a higher missed detection rate in small targets or complex scenarios. How to take into account detection accuracy and real-time performance while improving sensitivity to key characteristics of drowning (such as limb stiffness and sinking behavior) has become the core challenge for the implementation of current technology. The multi-model cascade method proposed in this paper aims to break through this technical bottleneck by integrating the advantages of YOLOv5 and Faster-RCNN and provides a more reliable solution for complex water scenes.

Deep learning neural network model training relies on a large number of training samples. However, drowning behavior usually occurs in water environments, and images of such scenarios are scarce in general datasets. At the same time, drowning behavior covers a variety of complex postures such as struggle and sinking, showing a high degree of behavioral diversity. The existing general target detection data sets, such as COCO, Pascal VOC, etc., do not contain such behavioral annotations and are not suitable for drowning detection tasks. In addition, drowning incidents themselves are relatively rare, and it is difficult to obtain relevant image and video data. The anti-drowning detection technology based on deep learning urgently needs high-quality data sets as support. If the data set problem is not solved in time, it may lead to delay in the implementation of the technology. Therefore, the collection of large-scale drowning data sets has become a difficult problem that needs to be overcome. To solve this problem, on the one hand, this experiment widely collects relevant pictures and data on multiple platforms online, and on the other hand, it goes deep into offline swimming pools to shoot materials and collect data through multiple channels.

3. Experimental Principles and Methods

3.1 YOLOv5 Image Recognition Algorithm

The YOLO algorithm is a real-time object detection algorithm. Its core idea is to transform object detection into a regression problem, using the entire picture as the input of the network, and through the neural network, the position of the bounding box and its category belonging to.

This paper uses the YOLOv5 algorithm to complete the drowning prevention detection work. The following is a detailed description of the network structure of YOLOv5.

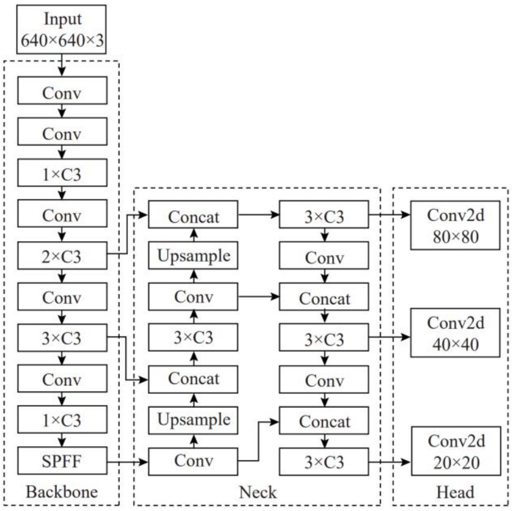


Figure 1. YOLOv5 Network Structure Diagram

As shown in Figure 1, the YOLOv5 neural network model consists of four parts: the input end, the backbone network, the neck network, and the head prediction layer. The input end uses Mosaic data enhancement to improve the problem of unbalanced data in the data center, small, medium and large target data, and uses adaptive anchor box calculation and adaptive picture scaling technology to reduce the amount of model calculation while improving algorithm performance. The backbone network is responsible for extracting the features of the input image, mainly composed of three modules: Conv, C3 and SPPF. The Conv convolution layer consists of convolution, Batch Normalization and SiLu activation layers for extracting features. The C3 module adaptively aggregates the previous feature map. The SPPF module obtains more comprehensive spatial information through the weighted fusion of global features and local features. The neck network is responsible for fusion of multi-scale features extracted by the backbone network and passing these features to the prediction layer. It adopts the FPN+PAN structure to fuse features of different scales through top-down and bottom-up paths. The head prediction end usually performs object detection on three feature maps of different scales to generate the final detection result.

3.2 Faster-RCNN Image Recognition Algorithm

Faster-RCNN is a two-stage object detection algorithm with higher accuracy than the single-stage object detection algorithm YOLO. The network structure is shown in Figure 2. The entire Faster-RCNN model can be divided into four modules: Conv layers, Region proposal Network, ROI pooling, Classification and Regression. After entering a picture, the Conv layers extracts image features through a set of conv + relu + pooling layers. The Region proposal Network uses the image features of the previous step as input, and uses the 3x3 convolution layer to convolutionize the feature map to generate intermediate features. One branch of the 1x1 convolution layer predicts whether each anchor contains the target (binary classification: foreground or background) to reshape and softmax operations on the classification results. At the same time, the other branch predicts the bounding box offset of each anchor point relative to the real box, and the two branches form a more accurate coordinate of the candidate area. The ROI pooling takes the area of interest output from the RPN network and the image features output from the feature extraction network as input.

After synthesising the two, a fixed-size area feature map is obtained and outputted to the subsequent fully connected network for classification. The Classification and Regression network inputs the regional feature map obtained from the previous layer, further extracts features through the fully connected layer, outputs the category to which the object belongs in the region of interest and the precise position of the object in the image. This layer classifies the image through softmax, and corrects the precise position of the object through border regression.

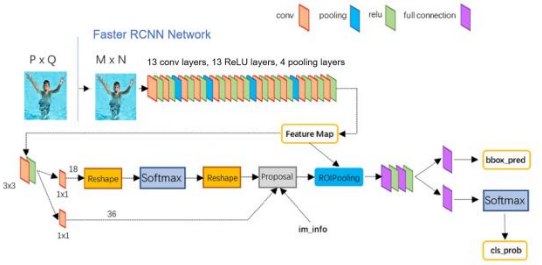


Figure 2. Faster-RCNN Network Structure Diagram

3.3 YOLOv5-FRNet

In the field of computer vision, Model Cascading, also known as Algorithm Ensemble, is a technical paradigm that synergizes multiple heterogeneous detection models to improve the overall performance of the system. Its core advantage lies in the comprehensive improvement of performance through the multi-model collaborative decision-making mechanism: in terms of detection capabilities, the complementary characteristics of different models can take into account both high accuracy and high recall, effectively covering the detection blind spots of a single model; in terms of robustness, the cross-verification mechanism of multi-models can significantly suppress false detection and missed detection caused by environmental interference. This integration paradigm is especially suitable for application scenarios that require strict detection reliability, and achieves a system-level performance jump of "1+1>2" through complementary advantages between models.

The YOLOv5-FRNet framework proposed in this paper adopts a confidence-driven cascade strategy. According to Figure 3, after inputting the photos or videos to be detected, independent detection of Faster R-CNN and YOLOv5 is first performed in parallel, and then the final detection results are generated through the confidence threshold screening and fusion mechanism (Confidence-based Fusion). The essence of this method belongs to Decision-level Ensemble. Its core idea is to use the differences in the confidence distribution of different models, and by setting a dynamic confidence threshold, retaining a high confidence prediction box and using Weighted-NMS for redundancy elimination, thereby achieving the optimal solution between accuracy and efficiency.

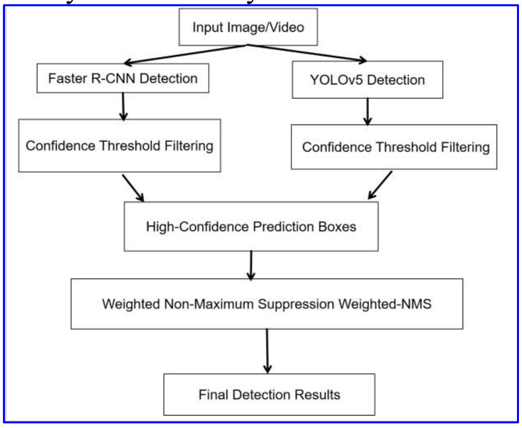

Input Image/Video
Faster R-CNN Detection
YOLOv5 Detection
Confidence Threshold Filtering
Confidence Threshold Filtering
High-Confidence Prediction Boxes
Weighted Non-Maximum Suppression Weighted-NMS
Final Detection Results

Figure 3. YOLOv5-FRNet Network Structure Diagram

4. Experimental Process

In order to verify whether the cascading algorithm is better than a single algorithm, this paper conducts experiments in the following 4 steps.

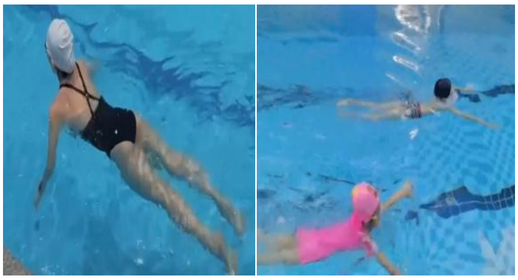
4.1 Experimental Preparation

At present, 8,518 data instances have been collected, of which 6,480 drowning images cover typical drowning postures such as struggle, head sinking, irregular limb movements, etc.; 2,038 swimming images include normal swimming movements such as freestyle and breaststroke and the steady water surface states (as shown in Figure 4 and Figure 5). The relevant pictures obtained can be used directly, and the videos are preprocessed through frame extraction before being used. Then, the resulting images are cropped and scaled uniformly, and the human posture states in the images are marked using the LabelImg tool.

The dataset is partitioned according to the ratio of training set: validation set: test sets 8:1:1. Of these, 6899 were used for training, 767 were used for validation, and 852 were used for testing.

(a)Drowning Example1 (b)Drowning Example 2

Figure 4. Example of Drowning Pictures in the Dataset



(a)swimming Example1 (b)swimming Example2

Figure 5. Example of Swimming Pictures in the Dataset

4.2 Model Training

The training process of the YOLOv5 model adopts dynamic optimization strategies and automated hyperparameter adjustment technology: the network structure is defined using yolov5s_swim.yaml, the input image is uniformly scaled to 640×640 pixels, and the data is annotated in COCO format. The data diversity is improved through Mosaic enhancement (probability 0.5), random flip, HSV color perturbation and multi-scale training (±50% scaling). The SGD optimizer (initial learning rate 0.01, momentum 0.9, weight decay 0.0005) is combined with linear learning rate decay and mixed precision training (AMP), with a batch size of 16, and the training is completed in 20 epochs (including a 3-epoch learning rate warm-up stage). AutoAnchor is introduced to automatically optimize the anchor box size and iteratively optimize the loss weight and data enhancement intensity based on the hyperparameter evolution (300-generation genetic optimization). Distributed Data Parallelism (DDP) acceleration is enabled during training, and the mean Average Precision at 0.5 (mAP@0.5) and the mean Average Precision at 0.5:0.95 (mAP@0.5:0.95) indicators are verified per epoch, and the optimal detection results are output through Non-Maximum Suppression (NMS) to ensure the balance between accuracy and efficiency of the model.

Faster R-CNN adopts a two-stage training strategy to balance training efficiency and model performance. First, the model is constructed based on the ResNet50 backbone network, the input image is uniformly adjusted to 600×600 pixels, and the VOC format data set is used and the training set and validation set are divided according to 2007_train.txt and 2007_val.txt. The training process is divided into two stages: freezing and thawing: In the first 10 epochs, the backbone network is frozen, and only the Region Proposal Network (RPN) and the detection heads are fine-tuned, with the batch size set to 4. The Adam optimizer (initial learning rate 1×10^{-4}, cosine annealing scheduled to 1×10^{-6}) is used. In the last 10 epochs, all network layers are thawed for end-to-end training, the batch size is adjusted to 2 and the learning rate is reset.

The RPN uses the combination of cross entropy loss and Smooth L1 loss for optimization, and the detection head uses class cross entropy loss and bounding box regression loss. The Average Precision at 0.5 (AP@0.5, conforming to the COCO standard) index is calculated on the validation set for every 5 epochs. Through the optimization of the training process through gradient clipping (threshold 10.0), FP16 hybrid precision training (GPU memory consumption <8GB), etc., the final model weights are saved in the logs directory. This strategy effectively balances training efficiency and model performance.

4.3 Model Testing and Evaluation

Through the methodology described in Section 4.2, the model weight files obtained from the training of YOLOv5 and Faster R-CNN can be used. These models are then tested on the test set respectively to evaluate the performance of each model on the drowning detection dataset. Additionally, the performance of YOLOv5 - FRNet on the test set is evaluated according to the process shown in Figure 3.

In order to compare the effects of the three methods, this experiment uses average accuracy, recall, and average accuracy mean as indicators to measure model performance, and the IOU threshold of the prediction box and the real box is set to 0.5. A good model performance is a sign that the average accuracy is higher.

The recall evaluation model finds incompleteness, which is the proportion of correctly predicted as positive samples in all data that are actually positive samples. Each category has the corresponding average accuracy. After calculating the accuracy and recall, the area under the accuracy-recall curve is obtained as the AP value, while the average value of the AP values of all categories is mAP. mAP is a common standard for target classification and border regression accuracy in statistical algorithms in the field of computer vision object detection. The larger the AP value, it means that the algorithm is more accurate in the target classification, the more accurate border regression, the fewer missing reports, the fewer false reports [11]. The calculation formulas of three indicators are as follows: Equations (1)-(3):

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (1)$$

$$\mathrm{AP} = \int (\mathrm{Precision})\, \mathrm{d}(\mathrm{Recall}) \quad (2)$$

$$\mathrm{mAP} = \frac{\sum_{\mathrm{N}}^{\mathrm{i}} \mathrm{AP_i}}{\mathrm{N}} \quad (3)$$

Where TP is the real positive sample number, FN is the falsely failed to detect positive sample number, and N is the total number of categories.

4.4 Analysis of Experimental Results

This experiment compares the performance of YOLOv5, Faster R-CNN, and the proposed YOLOv5 - FRNet cascade model for target detection tasks in drowning prevention scenarios. The key results are presented in Table 1:

Table 1. Comparison of Results of Different Algorithms

Algorithmic models	AP		R	mAP0.5	mAP0.5：0.95
	Drowning	Swimming			
YOLOv5	0.811	0.400	0.490	0.606	0.388
Faster-RCNN	0.890	0.639	0.593	0.764	0.480
YOLOv5-FRNet	0.890	0.658	0.603	0.774	0.603

Based on the data in Table 1, the following analysis is carried out:

Drowning Detection Accuracy: Both Faster R-CNN and YOLOv5 - FRNet achieved a Drowning - AP score of 0.890, which is a 9.7% improvement compared to YOLOv5 (0.811). This shows that the localization ability of two - stage detectors and cascaded models for drowning targets is significantly better than that of single - stage models. This is because the Region Proposal Network (RPN) of Faster R-CNN is sensitive to complex - pose targets, and cascaded models can retain high - quality prediction boxes through confidence screening.

Swimming Behavior Recognition Performance: In terms of Swimming - AP, YOLOv5 - FRNet (0.658) outperforms Faster R-CNN (0.639) by 3.0% and YOLOv5 (0.400) by 64.5% [(0.658 - 0.400) / 0.400 * 100%]. This verifies the adaptability of the cascade model to multi - category scenarios. Its performance improvement comes from the synergy between the high - recall characteristics of YOLOv5 and the fine - grained classification capabilities of Faster R-CNN.

Comprehensive detection efficiency (mAP): IoU=0.5 index: YOLOv5-FRNet (0.774) is optimized by 1.3% and 27.7% compared with Faster-RCNN (0.764) and YOLOv5 (0.606), respectively, reflecting the robustness of the cascade model for routine detection tasks; IoU=0.5:0.95 index: The cascade model (0.603) is 25.6% higher than Faster-RCNN (0.480), which is significantly better than YOLOv5 (0.388), proving its robustness advantage in multi-scale and multi-occlusion scenarios.

Recall rate: YOLOv5 - FRNet surpasses Faster R-CNN (0.593) and YOLOv5 (0.490) with a recall of 0.603. This indicates that the cascade strategy effectively reduces the missed - detection rate through Weighted Non - Maximum Suppression (Weighted - NMS), which is crucial for high - risk scenarios such as drowning monitoring.

5. Conclusion

This study focuses on water safety monitoring scenarios. The collected and annotated datasets are available to relevant researchers upon request. In addition, a confidence - driven YOLOv5 - FRNet cascade detection framework is proposed, and its significant advantages in drowning prevention detection tasks are verified through experiments. The cascading model YOLOv5 - FRNet has made a breakthrough in comprehensive detection efficiency, proving the robustness of the heterogeneous model cascade strategy for multi - scale and high - occlusion targets. Through the fusion mechanism of the confidence threshold and weighted NMS, the trade - off dilemma between precision and recall in traditional single models is effectively resolved. This method can be initially applied to swimming pools and other places for trial. In the future, we will attempt to develop a dynamic threshold adaptive mechanism based on meta - learning, further improve the deployment efficiency of edge devices, explore cross - scenario generalization training strategies, and enhance the model's adaptability to unseen water environments. This study provides a scalable technical framework for safety - critical visual detection tasks, which is of great significance for promoting the development of intelligent emergency rescue equipment.

References

[1] Maad S, Frdoos A, Ashwaq A, et al. Deep Learning and Vision-Based Early Drowning Detection. Information, 2023, 14 (1): 52-52.

[2] Yu Zhongyang, Wang Yajing. A Comparative Review of Drowning Monitoring Systems. Computer Knowledge and Technology, 2023, 19(13): 119-122+126.

[3] Liu T, He X, He L, et al. A video drowning detection device based on underwater computer vision. IET Image Processing, 2023, 17(6): 1905-1918.

[4] Wang Yun. Research on a Deep Learning-Based Early Warning and Monitoring Management System for Drowning Incidents. China New Technologies and Products, 2024, (01): 146-148.

[5] Jing Mingtao, Yu Teng, Feng Mengyao, Yang Guowei. Research on Drowning Detection in Swimming Pools Based on Improved Mask R-CNN. Journal of Qingdao University (Engineering & Technology Edition), 2021, 36(01): 1-7+21.

[6] Peng Ting, Shen Jinghu, Qiao Yu. Design of a Drowning Behavior Detection System in Swimming Pools Based on Improved Mask R-CNN. Transducer and Microsystem Technologies, 2021, 40(01): 94-97.

[7] Yu Zhongyang, Yang Wenhui. A Swimming Pool Drowning Detection Algorithm Based on Human Pose Estimation. Technology Innovation and Application, 2023, 13(23): 66-70+74.

[8] Chen Juan, Ge Bi, Chen Dongsheng. Research on a Reservoir Drowning Monitoring and Early Warning System. Computer Knowledge and Technology, 2024, 20(10): 5-7+14.

[9] Xie Jiangna, Liu Liqiu, Zhao Jialin, Wang Yiluan. Design of a Portable Swimming Pool Drowning Alarm and Physiological Health Assessment System. Electronic Engineering & Product World, 2023, 30(03): 19-23.

[10] Liu Xiangju, Shuai Tao, Jiang Shexiang. Drowning Personnel Detection Based on Improved YOLOv5. Journal of Shaanxi University of Technology (Natural Science Edition), 2024, 40(03): 35-43.

[11] Liu Songtao, Lü Hui, Liu Xiangling. Analysis of the Diagnostic Value of Artificial Intelligence Technology in Retinal Vein Occlusion. Recent Advances in Ophthalmology, 2025, 45(01): 46-49. DOI: 10.13389/ j. cnki. rao. 2025. 0009.