JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 05/06 pp62−73
Volume 4, Number 1, March 2023
DOI: 10.23919/JSC.2023.0006

Automatic Real-Time Detection of Infant Drowning Using YOLOv5 and Faster R-CNN Models Based on Video Surveillance

Qianen He, Zhiqiang Mei, Huisheng Zhang, and Xiuying Xu*

Abstract: Infant drowning has occurred frequently in swimming pools recent years, which motivates the research on automatic real-time detection of the accident. Unlike youths or adults, swimming infants are small in terms of size and motion range, and unable to send out distress signals in emergencies, which exerts negative effects on the detection of drowning. Aiming at this problem, a new step is initialized towards detecting infant drowning automatically and efficiently based on video surveillance. Diverse live-scene videos of infant swimming and drowning are collected from a variety of natatoriums and labeled as datasets. A part of the datasets is downscaled or enlarged to enhance generalization ability of the model. On this basis, advantages of Faster R-CNN and a series of YOLOv5 models are specifically explored to enable fast and accurate detection of infant drowning in real-world. Supervised learning experiments are carried out, model test results show that mean Average Precision (mAP) of either Faster R-CNN or YOLOv5s of the series of YOLOv5 can be over 89%; the former can process merely 6 frames of videos per second with the precision of only 62.04%, while the latter can reach an average speed of 75 frames/s with the precision of about 86.6%. The YOLOv5s eventually stands out as an optimal model for detecting infant drowning in view of comprehensive performance, which is of great application value to reduce the accidents in swimming pools.

Key words: infant drowning detection; YOLOv5; Faster R-CNN; video surveillance; supervised learning

1 Introduction

According to statistics, about 45% of drowning deaths worldwide involve children. In China, a country of over 1.4 billion population, drowning has become one of the three major causes of accidental injury and death to children[1]. With vigorous development of infant swimming industry, a growing number of parents encourage their infants to participate in swimming for the reason that swimming can improve children's physical fitness, coordination ability, and intelligence level, and promote neuro development as well. However, as the infant swimming prevails, safety-related problems have attracted widespread attention. Although infants are protected by swimming rings while swimming, accidents still occur mainly due to negligence of management staff and/or guardians. It is reported that a variety of infant drowning incidents have emerged in China recent years. These incidents are a devastating blow to the families and indicate that infant drowning remains a serious public safety problem. Hence, research and development of automatic real-time detection methods for infant drowning based on video surveillance have been urgent and of great significance.

Major contributions are briefly summarized as follows:

(1) The automatic real-time method for infant drowning detection is first proposed based on deep learning. Performances of the series of YOLOv5 (including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) and Faster R-CNN for infant drowning detection are tested and compared. And the YOLOv5s eventually stands out as the optimal

• Qianen He, Zhiqiang Mei, Huisheng Zhang, and Xiuying Xu are with the School of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China. E-mail: heqianen@tsinghua.org.cn; 862191244@qq.com; zhanghuisheng5610@163.com; xuxiuying@fzu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-09-27; revised: 2023-01-15; accepted: 2023-01-19

model with the superiority in comprehensive performance. It provides a new method and idea for infant drowning detection, which is of great application merit to reduce the accidents in swimming pools.

(2) Dataset of totally 7000 images is established for the first time, which contains 5000 images of normal swimming state and 2000 ones of drowning state. The former group of images is obtained by a consent to swimming pool owner or collected by social sharing software like Douyin, Xiaohongshu, etc. And the latter one is collected from on-site drowning videos reported online. Our dataset originates from 84 videos (49 videos depicting normal swimming and 35 videos depicting drowning), whose format is MP4 and average length is 38.8 seconds. And the number of infants observed and that of drowning ones in the videos are 133 and 35, respectively. The dataset utilized in this experiment has been made available on Kaggle, accessible through the following URL: https://www.kaggle.com/datasets/meizhiqiang/dataset-of-infants-swimming.

The remaining is organized as follows. Related work is described in Section 2. Framework of the method, the series of YOLOv5, and Faster R-CNN are introduced in Section 3. Experiments are conducted in Section 4. Experimental results are analyzed and discussed, and feasibility of the method is verified in Section 5. Conclusions are drawn in Section 6.

2 Related Work

At present, many scholars have carried out research on wearable sensor based drowning detections[2−5]. These studies mainly rely on the relationship between water pressure and time, the length of time for which the body stays underwater, abnormal changes in electrocardiogram, abnormal ripples to determine drowning or not. However, these methods are not suitable for the infant drowning detection due to the great possibility of the wearable device falling off and the strong discomfort caused by the device to infants.

Meanwhile, a lot of research has also been conducted on drowning detection based on images. Deep learning models, e.g., Mask R-CNN, transfer learning network, VIsual Background Extractor (VIBE), YOLOv4, Faster R-CNN, etc.[6−13], are used to detect drowning, and the accuracy rates of YOLOv4 and Faster R-CNN can reach 94.62% and 99.00%, respectively. Cha et al.[14, 15] adopted CNN and Faster R-CNN to detect concrete crack and steel corrosion, and mean Average Precision (mAP) can reach about 90.00%, which gives some ideas as to which models should be adopted. However, all these studies are aimed at adults, and are not tailored for the infant drowning detection, because drowning postures of adults and infants are quite different. The drowing postures of adults are often accompanied by violent physical movements, while the drowing postures of infants are usually like lying on one side or upside down in water, with swimming rings attached but without vigorous motions.

Currently, research on infant drowning automatic detection is still in its infancy. Gao[16] used SIFT algorithm and nine-block grid feature extraction algorithm to decide infant drowning or not, with an accuracy rate of 84%; however, the accuracy rate was only obtained by using an infant model instead of a real live one. Therefore, practical feasibility of the method remains a question. Scholars also determine infant drowning state based on abnormal electrocardiogram (ECG) changes detected by ECG sensors, or abnormal ripple changes detected by triaxial accelerometers[17−19]. However, these are only prevention systems for the drowning accidents in bathtubs at home rather than swimming pools. Hence, an automatic real-time method for infant drowning detection using YOLOv5 and Faster R-CNN based on video surveillance is proposed for the first time, where continuous frames of pictures are captured and analyzed in real time, and then decision is made on whether the infant is in a drowning state or not.

3 Method of Infant Drowning Detection

3.1 Framework of the method

As shown in Fig. 1, framework of the method includes three parts. Initially, a camera is used as image acquisition device to obtain scene information of the swimming pool. Subsequently, YOLOv5 and Faster R-CNN are trained using a dataset of 5600 images, during which the annotated data are input into the model for training, and Adaptive moment estimation (namely Adam) is employed as an optimizer of the model. Finally, as soon as the video is processed by the deep learning model, one of the two types of results will be determined and displayed: normal state or drowning state.

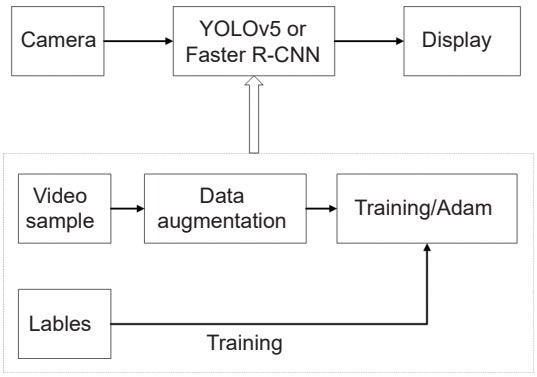


Fig. 1 Framework of the method.

3.2 Faster R-CNN and YOLOv5 models

3.2.1 Advantages of Faster R-CNN and YOLOv5

Currently, object detection models mainly fall into two categories: two-stage models and single-stage ones. Representative of the former is Faster R-CNN, where the detection is divided into two stages: region proposals are generated and their location coordinates are refined at the first stage, and then these region proposals are classified at the second stage[20]; while representative of the latter is the series of YOLOv5, where classification probabilities and position coordinates are directly generated without generating any region proposal.

The Faster R-CNN employs the region proposals in feature extraction, which enables finer detection and less missed detections, and boosts excellent performance in solving multi-scale and small-target detection problems[21]. The series of YOLOv5 performs well on detections of both small-size objects and overlapping ones while the processing speed can be so fast that requirement of real-time is easy to satisfy[22]. Therefore, the Faster R-CNN and YOLOv5 are selected as models on account of their potential merits in infant detection.

3.2.2 Structure of YOLOv5

Five models of the series of YOLOv5 (YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) are the same in terms of the structure, whereas their depth and width differ from one another[23]. Table 1 gives the metrics of performance for those, models tested on the COCO val2017 dataset. "Size" represents the size of the input image, mAP^{val}@0.5 represents the mAP of validation set when the Intersection Over Union (IOU) is set to 0.5, and Frames Per Second (FPS) denotes the detection speed.

The YOLOv5 structure contains four parts: input, backbone, neck, and prediction, as shown in Fig. 2. The input of the YOLOv5 employs the Mosaic data enhancement method to enrich the diversity of the drowning dataset via randomly scaling, cropping, and splicing the input images. The Spatial Pyramid Pooling (SPP) module of the "backbone" increases the receptive field of the convolution kernel, which can observe the information around the drowning infants, and improve the model's robustness to spatial layout and object deformation. The "neck" utilizes Feature Pyramid network (FPN) and Path Aggregation Network (PAN) in CSP2 to transfer semantic features and location features seperately in order to obtain more effective drowning information, and then the information is transferred to the prediction for

Table 1 Performance comparison of YOLOv5 five models[23].

Model	Size (pixel × pixel)	mAP^{val}@ 0.5 (%)	Number of parameters ($\times 10^6$)	FPS
YOLOv5n	640×640	45.7	1.9	222.2
YOLOv5s	640×640	56.8	7.2	60.6
YOLOv5m	640×640	64.1	21.2	20.4
YOLOv5l	640×640	67.3	46.5	9.2
YOLOv5x	640×640	68.9	86.7	4.8

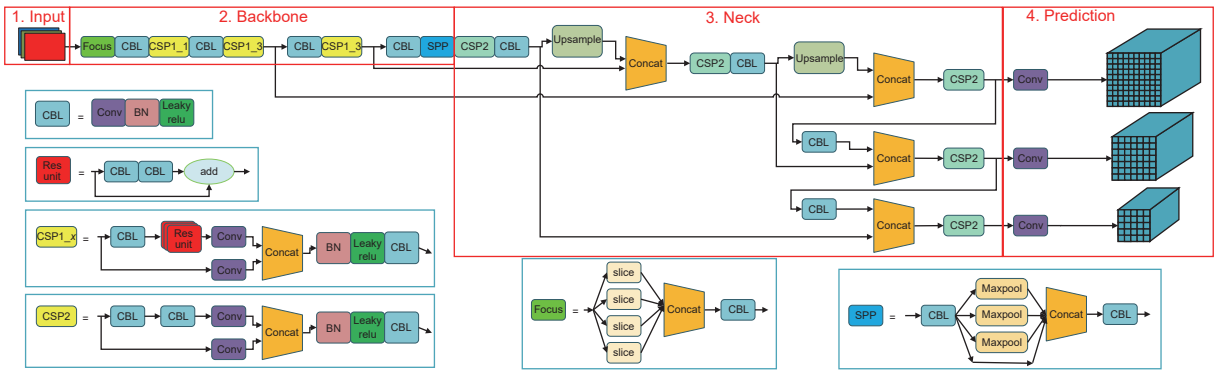


Fig. 2 YOLOv5 structure[24].

classification. The number of parameters of each convolution layer is shown in Table A1 in the Appendix.

3.2.3 Principle of the Faster R-CNN

The Faster R-CNN contains four parts: convolutional layers, Region Proposal Network (RPN), Region of Interest (RoI) pooling, and classifier[20], as shown in Fig. 3. Images are input to the convolutional layers for drowning feature extraction to obtain feature maps. Subsequently, the feature maps are passed to the RPN to generate a series of proposals for drowning or normal swimming, to which the classifier needs pay high attention in order to reduce the number of missed or false detections. Finally, these proposals and feature maps are transferred to the RoI pooling to make them to have uniform size, and are then passed to the classifier for classification. The number of parameters of each convolution layer is shown in Table A2 in the Appendix.

3.3 Supervised learning method

3.3.1 Data augmentation

Our original video dataset comprises of 7 videos depicting normal swimming and 5 depicting drowning. To improve the generalization ability of the deep learning model and increase the diversity of the dataset, we devote significant effort to data augmentation, employing various processing techniques, such as brightness adjustment, exposure processing, color difference processing, color rendering, mosaic processing, white noise processing, and scale transformation. As a result of our data augmentation, the equivalent number of normal swimming videos is expanded to 49, while the equivalent number of drowning videos is increased to 35. At the same time, our data augmentation efforts introduce variations in brightness, color, blur, size, angle, texture, and details of the swimming pool background, making the background more diverse. Additionally, the processing of the dataset introduces variations in skin color, head blurring, size, angle, texture, and details for the infants, rendering the infants more diverse. Consequently, the accuracy of our tested model is derived from the recognition of drowning or normal swimming actions performed by the infants.

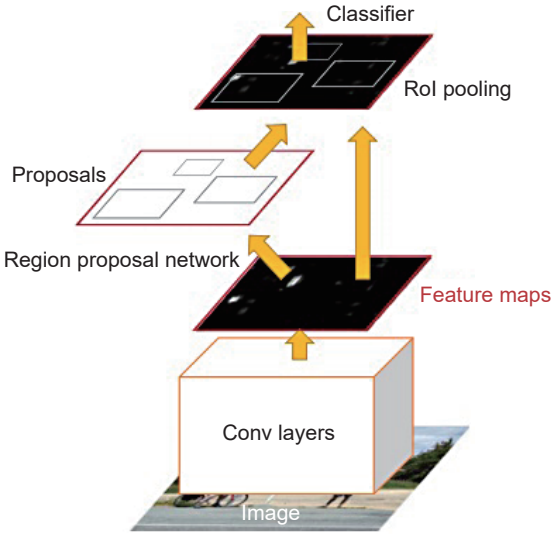


Fig. 3 Faster R-CNN structure[20].

3.3.2 Adam optimizer

Gradient descent is a method of finding the minimum value of a function. It first calculates the opposite direction vector of the gradient for current point of the function, and then an iterative search is performed with specified step size until the gradient approaches zero[25]. Adam, currently the most common gradient descent, not only has the merits of implementation simplicity, high computational efficiency, and low memory requirements, but also performs excellently in processing large-scale data and parameter optimization. As a result, Adam is used as an optimizer in the YOLOv5 and Faster R-CNN to replace the Stochastic Gradient Descent (SGD) optimizer. In the updating process of Adam, independent adaptive learning rates are designed for different parameters via estimating the first- and second-order moments of the gradient. For details of the process, please refer to Ref. [25].

4 Experiment

4.1 Annotation of the dataset

In this study, a total of 7000 images of infant swimming are collected as a dataset which is divided into training set, validation set, and test set in a ratio of approximately 3:1:1, as shown in Table 2. The training set and validation set are used to train the models and the test set is used to verify the performance of the models. These images are manually annotated by Make Sence Image Annotation Software to select target regions as accurately as possible, and the label format is set to YOLO.

Table 2 Distribution of the dataset.

Object	Number of normal images	Number of drowning images
Training set	3147	1053
Validation set	1039	361
Test set	1040	360

Based on observations of the collected images, the dataset is divided into two groups: normal swimming and drowning. The observed images, where upper body of any infant floats vertically on the water and is protected by a swimming ring, fall into the group of normal swimming, while those where any infant lies on its side or upside down in the water fall into the group of drowning.

4.2 Experimental environment and parameter configuration

Experiments are conducted on the Ubuntu18.04 operating system and the NVIDIA Ge Force GTX3060Ti GPU with 12 GB display memory. Softwares incloude Anaconda4.10.3, CUDA11.0GPU, Pytorch1.7.1, and Python 3.9. Tests are carried out on the models of Faster R-CNN and YOLOv5 (including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x). Besides, an attempt to conduct experiments on MindSpore is also made to further extend applications of the method.

In the training, the momentum is enabled and its value is set to 0.937, in order to speed up convergence of the loss function in the gradient descent method. Meanwhile, the weight decay is set to 0.0005; it is a prior-regularization factor which indicates the complexity of the model, and thus can adjust the influence of the model complexity on the loss function and prevent overfitting of the model during training. In addition, we also warm up the learning rate and momentum. At the beginning of the training, a small learning rate of 0.001 and a small momentum of 0.8 are initialized for the first 3 epochs to make the model gradually stabilize. And as soon as the model becomes stable, they change to the normal values, i.e., the learning rate of 0.01 and the momentum of 0.937, which enables the model converge faster and more effectively. Principal experimental parameters are shown in Table 3.

Table 3 Experimental parameter values.

Parameter	Value
Initial learning rate	0.01
Number of epochs	100
Batch size	16
Image size	640 pixel × 640 pixel
IOU training threshold	0.2
Optimizer	Adam
Momentum	0.937
Optimizer weight decay	0.0005
Number of warmup epochs	3.0
Warmup initial momentum	0.8
Warmup initial learning rate	0.001

4.3 Evaluation index

Performance of the model is evaluated by Precision rate (P), Recall rate (R), Average Precision (AP), and mAP. AP is obtained via calculating area of the region enclosed by the precision curve, the recall rate curve, and the axes. The mAP is obtained via accumulating the AP value of each category and dividing it by the number of categories. Notably, the mAP combinding the precision rate and recall rate can demonstrate the performance of the model more comprehensively. Therefore, the mAP is utilized as the main performance evaluation indicator among the aforementioned indicators. Equations for the four evaluation indexes are shown in the following:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{AP} = \int_0^1 P(r)\mathrm{d}r \tag{3}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{N} \mathrm{AP}_i}{N} \tag{4}$$

In Eqs. (1) and (2), True Positives (TP) represents the number of positive samples which are correctly predicted as positive, False Positives (FP) represents the number of negative samples which are incorrectly predicted as positive, False Negatives (FN) represents the number of positive samples which are incorrectly predicted as negative. And True Negatives (TN) represents the number of negative samples which are correctly predicted as negative. Meanings of TP, FP, TN, and FN are shown in Table 4. In Eq. (3), AP is obtained via integrating the smooth curve of the precision or recall rate. In Eq. (4), N is the number of categories of detection targets, and $\sum_{i=1}^{N} \mathrm{AP}_i$ represents the sum of the AP of each category.

5 Results and Discussion

5.1 Training results of the models

The dataset and the epoch of the experiment are kept the same for the series of YOLOv5 (including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) and the Faster R-CNN to enable equivalent comparison across the six models. Loss values on the validation set for the YOLOv5 series and the Faster

Table 4 Meanings of TP, FP, TN, and FN.

True result	Predicted result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

R-CNN are shown in Figs. 4a and 4b, respectively. It is worth noting that all the loss curves tend to be stable, which indicates that the training has already converged. Besides, the loss varying trends of the YOLOv5 series are consistently the same and have the merits of fast convergence and small convergence value, which indicates the YOLOv5 series are of excellent fitting effect and excellent performance of robustness. Particularly, faster convergence and smaller convergence value of the YOLOv5 series can be observed as compared to those of the Faster R-CNN.

5.2 Test results of the models

A set of 1400 images are used to test the trained models, and the confidence threshold is set to 0.5. Performance indicators for the test are summarized in Table 5. The mAP is used as the main performance indicator. And the FPS is used to evaluate the detection speed of the model, and the larger the value is, the faster the detection speed is; it is measured in the experiment where the NVIDIA GeForce GTX1650 graphics card is employed.

Obviously, the Faster R-CNN is not suitable for real-time infant drowning detection. On the one hand, Faster R-CNN has the structure of the two-stage, which intrinsically determines that the processing speed of the Fart R-CNN is considerably lower than those of the single-stage models of YOLOv5. As expected, test results show that the FPS of the YOLOv5s can reach 75, which is 69 higher than that of the Faster R-CNN. Notably, real-time detection can be realized as long as the FPS reaches 25−30 since high-resolution videos are usually of that speed. Hence, the Faster R-CNN of only 6 cannot meet the requirement of real-time while the YOLOv5s can.

On the other hand, the precision of the Faster R-CNN is only 62.04%, which implies that environmental background in the pool or normal state can be easily misidentified as drowning. Namely, if the Faster R-CNN were applied practically in drowning detection, false alarms would be issued frequently, which would cause chaos in the swimming pool. Besides, the Faster R-CNN is currently difficult to be integrated in embedded devices for drowning detection due to the excessive number of parameters. Therefore, the Faster R-CNN significantly lacks feasibility for detecting infant drowning in real time.

The YOLOv5s of the series of YOLOv5 eventually stands out as an optimal model for detecting infant drowning with the superiority of the comprehensive performance. Since its mAP reaches 89%, which outperforms other models of the YOLOv5 series, and the FPS of 75 can fully meet the real-time detection requirement. Furthermore, the YOLOv5s can be easily and conveniently integrated in embedded devices for infants drowning detection on account of the number of parameters of only 1.37×10^7.

5.3 Application cases of YOLOv5s

Load the real-time infant drowning detection model of YOLOv5s into the on-site video surveillance system, and then detection results are displayed, some of which are presented in Fig. 5. A red box labeled “Drowning” indicates the state of drowning, as shown in Fig. 5a; a greed box labeled “Normal” indicates the state of

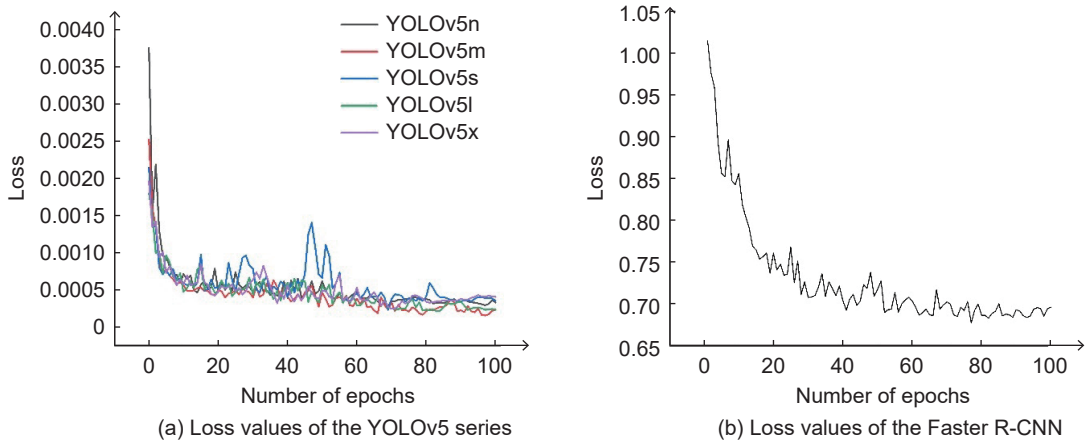


Fig. 4 Loss values of the YOLOv5 series and the Faster R-CNN on the validation set.

Table 5 Test results of the models.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	FPS	Number of parameters ($\times 10^6$)
YOLOv5n	90.50	71.30	88.60	167	3.65
YOLOv5s	86.60	76.30	89.00	75	13.70
YOLOv5m	97.20	68.30	88.10	32	40.20
YOLOv5l	95.80	64.70	85.20	20	88.50
YOLOv5x	97.90	66.60	87.40	11	165.00
Faster R-CNN	62.04	99.45	92.24	6	190.00

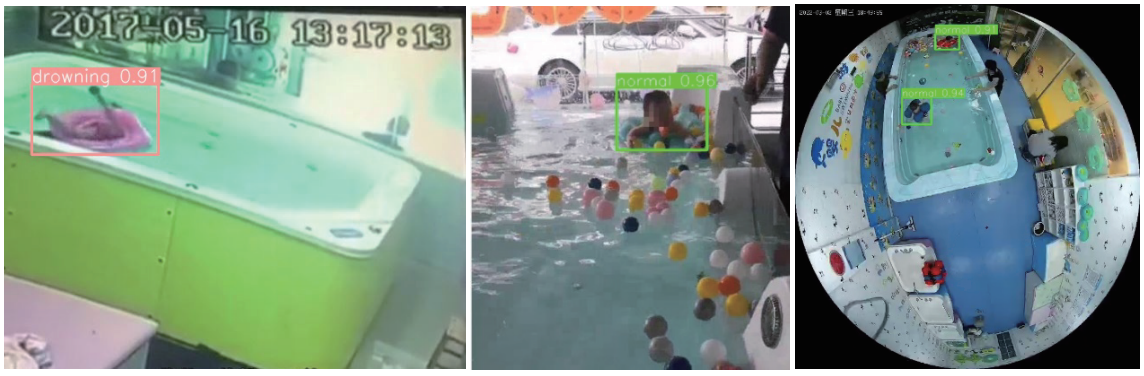
normal swimming, as shown in Figs. 5b and 5c. It can be seen that the YOLOv5s demonstrates excellent performance in detecting the states of infants, regardless of whether the infant is in a drowning state or not, and regardless of whether the camera is installed on the ceiling of the swimming pool or on the wall with approximately the same height as the water level. This takes into account the installation needs of cameras in practical applications and demonstrates the feasibility of the YOLOv5s.

However, the model may still have missed or false detections in the presence of some cases where the targets are extremely densely-distributed or immersed in complicated environment with toy disturbances. As shown in Figs. 6a and 6b, toys are misidentified as swimming infants. Additionally, there is situation presented in Fig. 6c where infants immersing in large area of toys are missed. The reason may be that the infant portrait is mixed with environmental noises due to the complicated and changing environment of the swimming pool, resulting in inaccurate features extracted by the YOLOv5s and thus leading to false or missed detections.

6 Conclusion

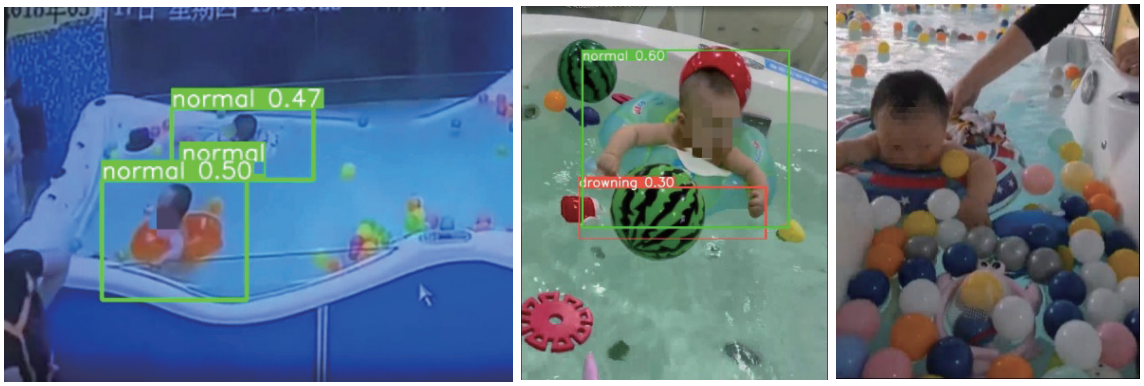
The first automatic real-time method using YOLOv5 and Fast-RCNN is first proposed for infant drowning detection based on video surveillance, which makes a major step to reduce the accident of infant drowning. Test results show that the mAP values of the Faster R-CNN and YOLOv5s reach 92.24% and 89%, respectively. However, the precision and FPS of the Faster R-CNN are only 62.04% and 6, respectively, resulting in a lack of practical feasibility when applied to real-time drowning detection. By contrast, the number of parameters and FPS of the YOLOv5s are 1.37×10^7 and 75, respectively, indicating that the YOLOv5s can be simply applied on embedded devices more conveniently and can fully meet the real-time requirement. Therefore, the YOLOv5s of the series of YOLOv5 eventually stands out as an optimal model. The mAP of the proposed method reaches 89% and the FPS is 45, which is higher than the method proposed by Beijing University of Technology in 2020, whose FPS is around 30 and accuracy is 84%[16].

Limitations of the work is that missed or false detections will rise when targets are extremely densely-distributed or immersed in complicated environment with toy disturbances. In the next step, we will optimize the YOLOv5s to accurately extract target features and enable the model to be better applied in infant drowning detection, aiming at providing a new



(a) (b) (c)

Fig. 5 Correct detections of the samples. (a) Detection of the drowning; (b) detection of the normal swimming when the camera is placed at a position close to the height of the swimming pool; and (c) detection of the normal swimming when the camera is placed on the ceiling.



(a) (b) (c)

Fig. 6 False detections of the samples. (a) and (b) False detections and (c) missed detection.

strategy to reduce or even eliminate the incidence of infant drowning.

Appendix

• Table A1 Number of parameters of each convolution layer in YOLOv5.

• Table A2 Number of parameters of each convolution layer in Faster R-CNN

Acknowledgment

This work was supported by the CAAI-Huawei MindSpore Open Fund and the General Program of Natural Science Foundation of Fujian Province, China (No. 2020J01473).

Data Reproducibility and Ethic Discussions

The data and experimental results are reproducible, and faces of children in the videos and images have been processed to hide their identities to meet the requirements of ethics and relevant laws and regulations. Datasets and related programs are available to anyone who is interested in the research by filing a request to the first author or corresponding author.

Table A1 Number of parameters of each convolution layer in YOLOv5.

Layer name	[number of input channels, number of output channels, size of convolution kernels, stride]×number of convolution kernels	Number of parameters
Conv1-x	[1, 1, 3, 1]×1	4
	[3, 4, 1×1, 1]×1	16
	[2, 1, 6×6, 2]×1	73
Conv2-x	[1, 1, 3, 1]×1	4
	[32, 4, 1×1, 1]×1	132
	[2, 1, 3×3, 2]×1	19
C3	[64, 32, 1×1, 1]×2	4160
	[64, 64, 1×1, 1]×1	4160
	[32, 32, 1×1, 1]×1	1056
	[32, 32, 3×3, 1]×1	9248
Conv3-x	[1, 1, 3, 1]×1	4
	[6, 4, 1×1, 1]×1	28
	[2, 1, 3×3, 2]×1	19
C3	[128, 64, 1×1, 1]×2	16 512
	[128, 128, 1×1, 1]×1	16 512
	[64, 64, 1×1, 1]×2	8320
	[64, 64, 3×3, 1]×2	73 856

(To be continued)

Table A1 Number of parameters of each convolution layer in YOLOv5. (Continued)

Layer name	[number of input channels, number of output channels, size of convolution kernels, stride]×number of convolution kernels	Number of parameters
Conv4-x	[1, 1, 3, 1]×1	4
	[128, 4, 1×1, 1]×1	516
	[2, 1, 3×3, 2]×1	19
C3	[256, 128, 1×1, 1]×2	65 792
	[256, 256, 1×1, 1]×1	65 792
	[128, 128, 1×1, 1]×3	49 536
	[128, 128, 3×3, 1]×3	442 752
Conv5-x	[1, 1, 3, 1]×1	3
	[256, 4, 1×1, 1]×1	1028
	[2, 1, 3×3, 2]×1	19
C3	[512, 256, 1×1, 1]×2	262 656
	[512, 512, 1×1, 1]×1	262 656
	[256, 256, 1×1, 1]×1	65 792
	[256, 256, 3×3, 1]×1	590 080
SPPF	[512, 256, 1×1, 1]×1	131 328
	[1024, 512, 1×1, 1]×1	524 800
Conv6	[512, 256, 1×1, 1]×1	131 328
C3	[512, 128, 1×1, 1]×2	131 328
	[256, 256, 1×1, 1]×1	65 792
	[128, 128, 1×1, 1]×1	16 512
	[128, 128, 3×3, 1]×1	147 584
Conv7	[256, 128, 1×1, 1]×1	32 896
C3	[256, 64, 1×1, 1]×2	32 896
	[128, 128, 1×1, 1]×1	16 512
	[64, 64, 1×1, 1]×1	4160
	[64, 64, 3×3, 1]×1	36 928
Conv8	[128, 64, 1×1, 1]×1	8256
C3	[128, 32, 1×1, 1]×1	4128
	[128, 32, 1×1, 1]×1	4128
	[64, 64, 1×1, 1]×1	4160
	[32, 32, 1×1, 1]×1	1056
	[32, 32, 3×3, 1]×1	9248
Conv9	[64, 64, 3×3, 2]×1	36 928
C3	[320, 64, 1×1, 1]×2	41 088
	[128, 128, 1×1, 1]×1	16 512
	[64, 64, 1×1, 1]×1	4160
	[64, 64, 3×3, 1]×1	36 928
Conv10	[128, 128, 3×3, 2]×1	147 584
C3	[640, 128, 1×1, 1]×1	82 048
	[640, 128, 1×1, 1]×1	82 048
	[256, 256, 1×1, 1]×1	65 792
	[128, 128, 1×1, 1]×1	16 512
	[128, 128, 3×3, 1]×1	147 584
Conv11	[256, 256, 3×3, 2]×1	590 080

(To be continued)

Table A1 Number of parameters of each convolution layer in YOLOv5. (Continued)

Layer name	[number of input channels, number of output channels, size of convolution kernels, stride]×number of convolution kernels	Number of parameters
C3	[512, 256, 1×1, 1]×2	262 656
	[512, 512, 1×1, 1]×1	262 656
	[256, 256, 1×1, 1]×1	65 792
	[128, 128, 3×3, 1]×1	147 584
Detect	[64, 18, 1×1, 1]×1	1170
	[128, 18, 1×1, 1]×1	2322
	[256, 18, 1×1, 1]×1	4626
	[512, 18, 1×1, 1]×1	9234

Table A2 Number of parameters of each convolution layer in Faster R-CNN

Layer name		[number of input channels, number of output channels, size of convolution kernels, stride]×number of convolution kernels	Number of parameters
Conv1		[3, 64, 7×7, 2]×1	9472
Extractor	Conv2-x	[64,64,1×1,1]×1	
		[64,64,3×3,1]×1	57 728
		[64,256,1×1,1]×1	
		[256, 64, 1×1, 1]×2	
		[64, 64, 3×3, 1]×2	140 032
		[64, 256, 1×1, 1]×2	
	Conv3-x	[256, 128, 1×1, 2]×1	
		[128, 128, 3×3, 1]×1	213 504
		[128, 256, 1×1, 1]×1	
		[512, 128, 1×1, 1]×3	
		[128, 128, 3×3, 1]×3	2 066 304
		[128, 256, 1×1, 1]×3	
	Conv4-x	[512, 256, 1×1, 2]×3	
		[256, 256, 3×3, 1]×3	984 576
		[256, 1024, 1×1, 1]×3	
		[1024, 256, 1×1, 1]×5	
		[256, 256, 3×3, 1]×5	5 578 240
		[256, 1024, 1×1, 1]×5	
RPN	Conv5	[2048, 256, 3×3, 1]×1	4 718 848
	Conv6	[256, 18, 1×1, 1]×1	4626
	Conv7	[256, 36, 1×1, 1]×1	9252
ROI	Conv8-x	[1024, 512, 1×1, 2]×1	
		[512, 512, 3×3, 1]×1	3 935 232
		[512, 2048, 1×1, 1]×1	
	Conv9-x	[2048, 512, 1×1, 1]×2	
		[512, 512, 3×3, 1]×2	8 919 040
		[512, 2048, 1×1, 1]×2	

References

[1] M. N. Dai, Y. Xi, W. Q. Yin, Z. M. Chen, Z. Q. Feng, and C. H. Tang, Incidence, mortality and trends of drowning among children aged 0–14 years in China, 1990–2019, *Chin. J. School Health*, vol. 43, no. 2, pp. 256–259&264, 2022.

[2] F. Wang, Y. B. Ai, and W. D. Zhang, Detection of early dangerous state in deep water of indoor swimming pool based on surveillance video, *Signal*, *Image and Video Processing*, vol. 16, no. 1, pp. 29–37, 2022.

[3] A. Kulkarni, K. Lakhani, and S. Lokhande, A sensor based low cost drowning detection system for human life safety, in *Proc. 5th Int. Conf. Reliability*, *Infocom Technologies and Optimization* (*Trends and Future Directions*) (*ICRITO*), Noida, India, 2016, pp. 301–306.

[4] F. Dehbashi, N. Ahmed, M. Mehra, J. Wang, and O. Abari, SwimTrack: Drowning detection using RFID, in *Proc. ACM SIGCOMM 2019 Conf. Posters and Demos*, Beijing, China, 2019, pp. 161–162.
[5] A. Jose and G. Udupa, Gantry robot system for preventing drowning accidents in swimming pools, *Mater. Today: Proc*., doi: 10.1016/j.matpr.2020.10388.
[6] Y. Y. Wang, Video monitoring system design and implementation of the underwater of pool, (in Chinese), Master dissertation, Beijing University of Technology, Beijing, China, 2014.
[7] J. B. Hou and B. G. Li, Swimming target detection and tracking technology in video image processing, *Microprocessors and Microsystems*, vol. 80, p. 103535, 2021.
[8] A. Alotaibi, Automated and intelligent system for monitoring swimming pool safety based on the IoT and transfer learning, *Electronics*, vol. 9, no. 12, p. 2082, 2020.
[9] A. Claesson, S. Schierbeck, J. Hollenberg, S. Forsberg, P. Nordberg, M. Ringh, M. Olausson, A. Jansson, and A. Nord, The use of drones and a machine-learning model for recognition of simulated drowning victims-A feasibility study, *Resuscitation*, vol. 156, pp. 196–201, 2020.
[10] M. A. Hayat, G. T. Yang, A. Iqbal, A. Saleem, A. Hussain, and M. Mateen, The swimmers motion detection using improved VIBE algorithm, in *Proc. Int. Conf. Robotics and Automation in Industry* (*ICRAI*), Rawalpindi, Pakistan, 2019, pp. 1–6.
[11] A. I. N. Alshbatat, S. Alhameli, S. Almazrouei, S. Alhameli, and W. Almara, Automated vision-based surveillance system to detect drowning incidents in swimming pools, in *Proc. Advances in Science and Engineering Technology International Conf.* (*ASET*), Dubai, United Arab Emirates, 2020, pp. 1–5.
[12] F. Lei, H. Y. Zhu, F. F. Tang, and X. Y. Wang, Drowning behavior detection in swimming pool based on deep learning, *Signal Image and Video Processing*, vol. 16, no. 6, pp. 1683–1690, 2022.
[13] P. Pavithra, S. Nandini, A. Nanthana, N. T. Aslam, and P. Praveen Kumar, Video based drowning detection system, in *Proc. Int. Conf. Design Innovations for 3Cs Compute Communicate Control* (*ICDI3C*), Bangalore, India, 2021, pp. 203–206.
[14] Y. J. Cha, W. Choi, and O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017.
[15] Y. J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types, *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 9, pp. 731–747, 2018.
[16] Y. N. Gao, Research on the method of drowning monitoring in infant swimming pool, (in Chinese), Master dissertation, Beijing University of Technology, Beijing, China, 2020.
[17] Y. Deng and T. H. Zhou, Sensor-based self-rescue alarm system for the prevention of child drowning, in *Proc. 13th Int. Conf. Intelligent Human-Machine Systems and Cybernetics* (*IHMSC*), Hangzhou, China, 2021, pp. 168–171.
[18] Y. Nishida, K. Hiratsuka, and H. Mizoguchi, Prototype of infant drowning prevention system at home with wireless accelerometer, in *Proc. SENSORS,* Atlanta, GA, USA, 2007, pp. 1209–1212.
[19] K. Hiratsuka, Y. Nishida, and H. Mizoguchi, Infant drowning prevention system with wireless accelerometer—Evaluation of optimum floating body shape for home-use, in *Proc. SENSORS*, Lecce, Italy, 2008, pp. 1218–1221.
[20] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
[21] B. Liu, W. C. Zhao, and Q. Q. Sun, Study of object detection based on Faster R-CNN, in *Proc. Chinese Automation Congress* (*CAC*), Ji'nan, China, 2017, pp. 6233–6236.
[22] S. Luo, J. Yu, Y. J. Xi, and X. Liao, Aircraft target detection in remote sensing images based on improved YOLOv5, *IEEE Access*, vol. 10, pp. 5184–5192, 2022.
[23] Ultralytics, https://github.com/ultralytics/yolov5/tree/v5.0, 2022.
[24] B. Li, M. M. Fu, and Q. Li, Runway crack detection based on YOLOV5, in *Proc. IEEE 3rd International Conference on Civil Aviation Safety and Information Technology* (*ICCASIT*), Changsha, China, 2021, pp. 1252–1255.
[25] S. Bock and M. Weiß, A proof of local convergence for the Adam optimizer, in *Proc. Int. Joint Conf. Neural Networks* (*IJCNN*), Budapest, Hungary, 2019, pp. 1–8.

Qianen He received the BEng degree in measurement and control technology and instruments from Harbin Institute of Technology, China in 2008, and the PhD degree in instrument science and technology from Tsinghua University, China in 2013. From 2013 to 2015, he was a postdoctoral researcher at the Department of Precision Instruments, Tsinghua University. Since 2015, he has been engaged in research and teaching in optical and electronic information engineering at the School of Physics and Information Engineering, Fuzhou University. His research interests include computer-based measurement and control, signal processing, and system modeling.

Zhiqiang Mei received the BEng degree in material physics from Fujian Normal University, China in 2021. He is currently a master student in integrated circuit at the School of Physics and Information Engineering, Fuzhou University, China. His research interests include machine learning, as well as data mining and AI/data for social good.

Xiuying Xu received the BEng degree in pharmacy engineering in 2002 and MEng degree in communication & information system in 2004 both from Nanjing University of Science and Technology, China. Since 2004, she has been engaged in research and teaching in electronic information engineering at the School of Physics and Information Engineering, Fuzhou University. Her research interests include signal processing and pattern recognition.

Huisheng Zhang received the BEng degree in applied physics from Hangzhou Normal University, China in 2021. He is currently a master student in circuits and systems at the School of Physics and Information Engineering, Fuzhou University, China. His research interests include machine learning, as well as data mining and AI/data for social good.