applied sciences

MDPI

Article

CTDR-Net: Channel-Time Dense Residual Network for Detecting Crew Overboard Behavior

Zhengbao Li, Jie Gao, Kai Ma, Zewei Wu and Libin Du *

College of Ocean Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China; lizhengbao@sdust.edu.cn (Z.L.); 17805424570@163.com (J.G.); michaellzz@163.com (K.M.); w67776756@126.com (Z.W.)
* Correspondence: dulibinhit@163.com

Abstract: The efficient detection of crew overboard behavior has become an important element in enhancing the ability to respond to marine disasters. It remains challenging due to (1) the lack of effective features making feature extraction difficult and recognition accuracy low and (2) the insufficient computing power resulting in the poor real-time performance of existing algorithms. In this paper, we propose a Channel-Time Dense Residual Network (CTDR-Net) for detecting crew overboard behavior, including a Dense Residual Network (DR-Net) and a Channel-Time Attention Mechanism (CTAM). The DR-Net is proposed to extract features, which employs the convolutional splitting method to improve the extraction ability of sparse features and reduce the number of network parameters. The CTAM is used to enhance the expression ability of channel feature information, and can increase the accuracy of behavior detection more effectively. We use the LeakyReLU activation function to improve the nonlinear modeling ability of the network, which can further enhance the network's generalization ability. The experiments show that our method has an accuracy of 96.9%, striking a good balance between accuracy and real-time performance.

Keywords: crew overboard; CTDR-Net; DR-Net; CTAM; good balance

Citation: Li, Z.; Gao, J.; Ma, K.; Wu, Z.; Du, L. CTDR-Net: Channel-Time Dense Residual Network for Detecting Crew Overboard Behavior. *Appl. Sci.* **2024**, *14*, 986. https://doi.org/10.3390/app14030986

Academic Editor: Eui-Nam Huh

Received: 29 December 2023
Revised: 21 January 2024
Accepted: 23 January 2024
Published: 24 January 2024

1. Introduction

Marine disasters, such as wind surges, tsunamis, waves, ocean earthquakes, and underwater volcanoes, are occurring more frequently with global climate change in recent years, which has seriously affected maritime navigation safety and brought significant risks to maritime economic activities [1]. Crew overboard accidents occur frequently due to the impact of marine disasters and limitations in the working environment at sea, which cause substantial loss of life and property. A rapid and accurate method for detecting crew overboard can reduce warning time, improve rescue response efficiency, and reduce loss of life and property. The method has become an important aspect in improving the emergency response capabilities of marine disasters and the rescue efficiency of maritime accidents.

Traditional overboard detection systems utilize various sensing devices such as pressure sensors, infrared sensors, and sound sensors to detect overboard incidents. Sevin et al. designed W-MEDS, a Wireless Sensor Network (WSN)-based system for swiftly detecting and locating individuals in man-overboard emergencies on ships [2]. W-MEDS uses real-time sensing by WSN nodes (temperature, humidity, acceleration) to trigger alarms and initiate rescue procedures through a control and discovery system. In order to enhance the notification speed and the accuracy of locating individuals who have fallen overboard, the feasibility of using wireless transmitters was investigated [3]. After conducting a literature review, the choice of technology fell on LoRa. Field tests were conducted in a maritime environment, revealing challenges in employing wireless technology for detecting and tracking overboard individuals on large cruise ships. These challenges are primarily attributed to the physical limitations of radio wave propagation in water. Sheu et al. proposed a real-time MOB alert, GPS tracking, and monitoring system [4]. The

system includes wearable sensors, remote LoRa access points, physical electronic fences, and three MOB detection methods: real-time notification via wearables, virtual electronic fence monitoring based on ship size, and instant notification triggered by a surrounding electronic fence. Laboratory and sea tests demonstrated the system's ability for quick MOB detection and notification, showcasing the ship's real-time MOB detection and prompt rescue capabilities. Yan et al. proposes a joint time-frequency domain processing method for the detection of man-overboard signals [5]. It utilizes empirical modal decomposition to separate and detect multi-component signals for the purpose of effectively detecting men overboard. The traditional overboard detection systems focus on deploying sensing devices to activate the overboard alarm. The forms of data collection and transmission are easily affected by and interfered with by external environments. These factors result in high system maintenance costs [6].

Crew overboard detection technology is progressively evolving towards automation and intelligence, driven by the advancement of mobile communication technology and artificial intelligence [7]. The intelligent detection of men overboard with computer vision and image processing techniques is a current research focus. The PSPNet (Pyramid Scene Parsing Network) deep learning model is proposed to achieve pedestrian overboard detection [8]. It uses the principle of motion detection to calculate the distance between the pedestrian center of mass and the contour of the lake, detecting overboard incidents by judging the distance. The YOLO-WA (you only look once-water area) algorithm is proposed for target detection in waterside environments [9]. It selects activation functions, introduces attention mechanisms, and replaces loss functions to improve the accuracy of human recognition in aquatic settings. Yang introduces a man overboard detection algorithm applicable to the surrounding waters of pump ships [10]. It employs an enhanced YOLOv4 for human detection, utilizes DeeplabV3+ for water surface segmentation, and determines whether the target is in contact with the water surface. Zhang introduces a YOLO-based target detection algorithm using infrared images [11]. It streamlines the backbone network and incorporates an attention mechanism to enhance the success rate and efficiency of detecting small targets. The image-based overboard detection methods overlook the influence of overboard behavior coherence on detection accuracy. Meanwhile, factors such as light scattering and water surface ripples can also reduce the accuracy of overboard detection, leading to false and missed detections.

Researchers utilize surveillance video frame sequences to design crew overboard behavior detection algorithms. Traditional models for behavior detection rely on manually designed features with poor generalization. Deep learning-based behavior detection models can effectively capture abstract features by learning massive amounts of data [12], which improves the generalization performance of the models in diverse scenarios [13]. A two-stream convolutional neural network that acquires features from RGB image and optical flow is proposed for behavior detection [14]. A Temporal Segment Network (TSN) is introduced to address the deficiency in long-term temporal modeling in dual-stream networks [15]. The dual-stream-based detection method can effectively extract the temporal and spatial features of the behavior. However, this method has difficulties in extracting the optical flow information of the video and requires a large amount of resources, which cannot meet the real-time requirements of crew overboard detection. Tran et al. utilized three-dimensional convolutional (C3D) neural network to directly extract the temporal and spatial features of the behavior [16]. Carreira and Zisserman proposed an Inflated 3D Convolutional Neural Network (I3D) [17], which was pre-trained on large-scale datasets and they fine-tuned the model to adapt to a specific behavioral detection task. Hara et al. combined 3D convolution with ResNet to increase the network depth and improve the accuracy of behavior detection [18]. Qiu et al. proposed a Pseudo-3D Residual Network (P3D Resnet) to reduce the computational load of 3D convolutional networks [19]. These methods can effectively solve the temporal–spatial modeling problem and improve the accuracy of men overboard behavior detection, but fail to effectively distinguish behavioral features and background information. Hu et al. introduced a channel attention model

known as Squeeze-and-Excitation Networks (SENet) to augment the network's attention to different channel features [20]. Woo et al. combined channel attention with spatial attention to enhance the model's focus on crucial channels and regions [21].

In the video-based monitoring system for crew behavior, the surveillance camera is positioned at the edge of the upper deck of the ship. It oversees the peripheral railing area of the ship from a high monitoring perspective. Inclement weather such as storms, heavy surf, and fog can reduce the clarity of video images. Meanwhile, the video image information acquired by the camera mostly contains the back and side of the crew, which has fewer usable features as the crew pose is obscured by the ship and the equipment on board. The pose information is completely lost due to the obstruction of the ship's hull when the crew falls over the side of the ship. The limited intelligent computing resources of ships exacerbate the difficulty of designing high-performance algorithms. From the above analysis, we can conclude that the crew overboard behavior has fewer available features and shorter durations. These characteristics increase the difficulty of abnormal behavior feature extraction and reduce the accuracy of detection. The proposed behavior detection algorithms have an imbalance in accuracy in real-time, which makes it difficult to meet the demand for crew overboard behavioral detection algorithms.

In this paper, we propose a crew overboard behavior detection network based on a dense residual block, tailored to the characteristics of crew overboard behavior. We use a sparse sampling method to extract video frame sequences, which form the input of the model and can improve the computational efficiency of the model. DR-Net is constructed as a backbone feature network to extract the behavioral features and convolutional splitting is used to reduce the number of network parameters. In DR-Net, dense blocks are utilized to fully extract the behavioral features in the shallow network and residual blocks are employed to extract the deep features to reduce overfitting. In addition, we introduce CTAM to efficiently extract key behavioral features and adopt the LeakyReLU activation function to enhance the robustness and generalization of the model, which can improve the detection performance of the network.

2. Overview of CTDR-Net

Our Channel-Time Dense Residual Network (CTDR-Net) can be seen in Figure 1. The process of overboard behavior usually lasts for 2~4 s, and a camera with a frame rate of 25 fps can collect approximately 50~100 frames of images. We use a sparse sampling method with a two-frame interval to extract video frame sequences, which are input into the model for recognition. This method can reduce the number of input images and the computational complexity of the model. We set the size of the model input to $3 \times 32 \times 224 \times 224$, where 32 represents the input frame sequence. According to the analysis of the duration and characteristics of the overboard behavior, it can be concluded that 32 frames of images can contain more than half of the overboard behavior information.

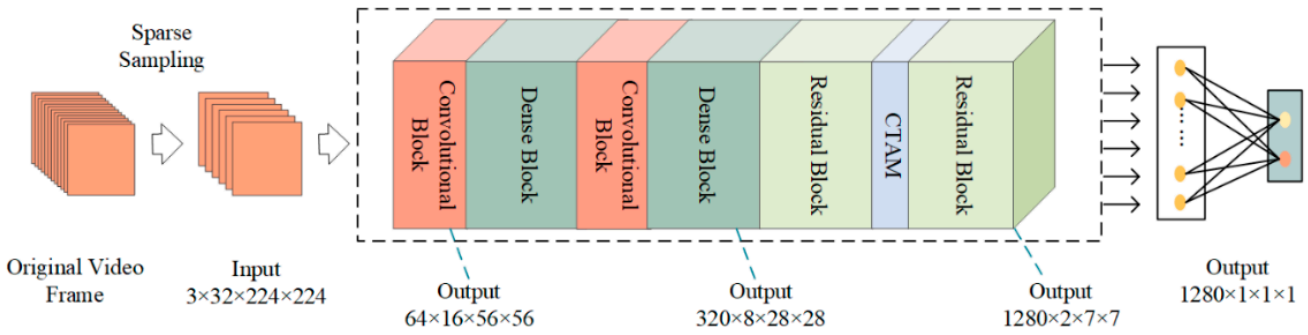


Figure 1. Illustration of our CTDR-Net.

When crew members are overboard, they are often obscured, which increases the difficulty of feature extraction. DR-Net is designed to extract features from the input

image. It uses $7 \times 7 \times 3$ convolution and $3 \times 3 \times 3$ maximum pooling to extract the base features and reduce the dimensionality of the input data. DR-Net employs two dense blocks and transition convolutional layers to improve the feature extraction capability and compress the features to reduce the number of parameters. Two residual blocks are adopted to enhance the deep feature extraction capability, which can reduce the risk of network overfitting.

We introduce the CTAM in the middle of the two residual blocks to filter out the background noise and increase the key feature extraction capability. The fully connected layer combined with softmax activation function is used to realize crew overboard behavior detection. We adopt the LeakyReLU activation function to enhance the robustness and generalization ability of the network and the cross-entropy loss function and back propagation algorithm to optimize the model parameters, which can improve the detection accuracy of crew overboard behavior.

3. DR-Net and CTAM

3.1. DR-Net

The video images obtained by the onboard video monitoring system contain less frontal information of crew members, which is limited by the deployment location of the cameras. Crew members work in complex environments and their limbs are often obscured by equipment and cargo, which leads to difficulties for the camera in capturing complete body features. Severe weather reduces the clarity of monitoring images and further increases the difficulty of feature recognition. These factors result in fewer crew features obtained through surveillance videos, making feature extraction difficult and detection accuracy low. The lack of information can be compensated for by increasing the number of layers of the feature extraction network, but it can easily lead to the problem of overfitting and poor generalization ability.

In this paper, a new feature extraction network, Dense Residual Network (DR-Net), is constructed by combining the dense block and residual block to improve the feature extraction capability of crew transgression behavior.

3.1.1. Dense Blocks Design

The output of the previous layer is only used as the input of the next layer in traditional convolutional neural networks, which constitutes less feature reuse and is not applicable to the detection applications with sparse features in this paper. The dense block structure consists of multiple dense layers, and the input of each dense layer is composed of the outputs of all preceding layers. In this structure, shallow features can participate in deep feature extraction, which avoids information sparsity and increases the efficiency of feature utilization.

Considering the characteristics and computational complexity of crew overboard behavior feature extraction, two dense blocks are added to the DR-Net to extract shallow image features in the video frame sequence. Figure 2 shows a sketch of the dense blocks, where each dense block contains six dense layers, with a transition layer introduced between the two dense blocks.

The new feature map is formed by concatenating the original input of the dense block with the output of all the previous dense layers, and serves as the input for the current dense layer. Each dense layer produces k output feature maps; the size of the input and output feature maps remains consistent. As we can see from Figure 2, the l^{th} dense layer has $k_0 + k \times (l-1), l = 1, \ldots, 6$ input feature maps, where k_0 is the number of original input channels in the dense block. In each dense layer, the number of the input feature maps is compressed to $4 \times k$ through $1 \times 1 \times 1$ convolution operation, where $k = 32$. This compression operation reduces the number of input features in dense layers and reduces the computational complexity of the network. K pseudo-convolutions of size $1 \times 3 \times 3$ and $3 \times 1 \times 1$ are utilized to replace the traditional convolution of size $3 \times 3 \times 3$, which can improve the real-time detection performance. The $(l+1)^{\text{th}}$ dense layer has

$k_0 + k \times l$, $l = 1, \ldots, 6$ input feature maps, which are composed of the outputs of all previous layers. After feature extraction in six dense layers, $6 \times k$ feature maps are added to each dense block.

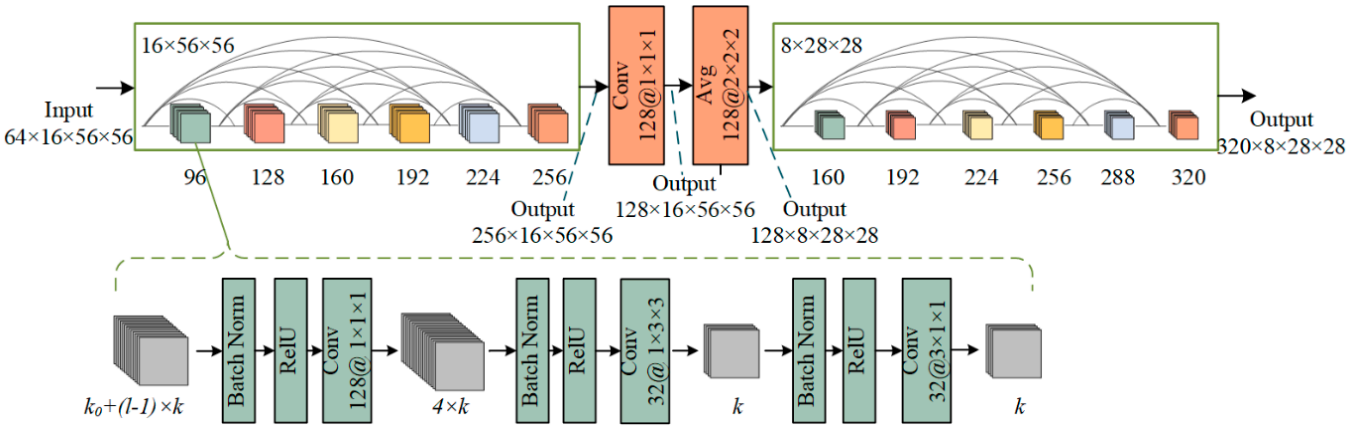


Figure 2. Dense blocks.

The transition layer connecting two adjacent dense blocks is composed of a $1 \times 1 \times 1$ convolutional layer and a $2 \times 2 \times 2$ average pooling layer. The $1 \times 1 \times 1$ convolutional layer reduces the number of channels of the dense block output feature maps from 256 to 128, which reduces the computational complexity and maintains feature extraction accuracy. The $2 \times 2 \times 2$ pooling layer reduces the size of the output feature map from 56×56 to 28×28, which preserves important information and reduces redundancy.

3.1.2. Residual Blocks Design

We use dense blocks to enhance information transmission and feature reuse, but the increase in the number of layers in the feature extraction network can easily lead to overfitting. DR-Net introduces two residual blocks after two dense blocks to enhance the deep feature extraction capability and reduce the risk of overfitting.

Figure 3 shows a sketch of the residual block in DR-Net. The downsampling residual module reduces the size of the input feature map to half of the original size and doubles the channel number, while the residual module does not change the size of the input feature map or the channel number. The first residual block consists of 1 downsampling residual module and 10 residual modules. The second residual block consists of one downsampling residual module and two residual modules. The downsampling residual module uses $1 \times 1 \times 1$ convolution to fuse the input feature maps, which keeps the number of feature maps unchanged and preserves the input feature information to the maximum extent. The $3 \times 3 \times 3$ convolution is replaced by $1 \times 3 \times 3$ and $3 \times 1 \times 1$ pseudo-convolution with a step size of two, which reduces the size of the feature maps to half of the original size and improves the real-time performance. The $1 \times 1 \times 1$ convolutions (n = 640, 1280) separately extract features from the original input feature maps (with stride = 2) and the half-sized feature maps (with stride = 1). The output feature maps of these two operations are added together as the output of the downsampling residual module.

The residual module is similar in structure to the downsampling residual module, except that it does not have a downsampling operation, because its input feature map and output feature map have the same size.

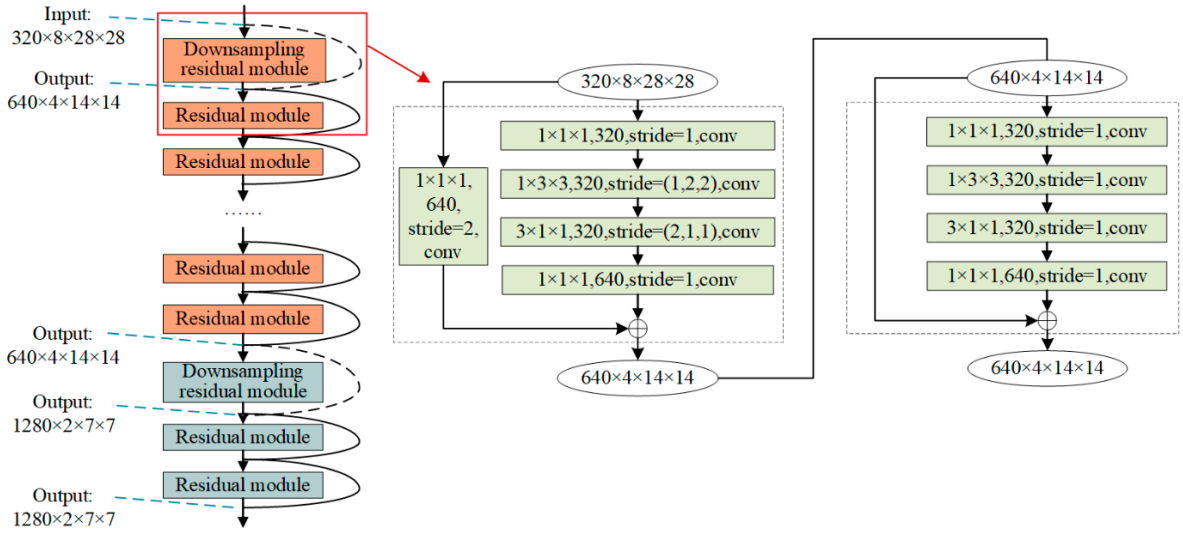


Figure 3. Residual blocks.

3.2. Channel-Time Attention Mechanism (CTAM)

An attention mechanism is introduced between two residual blocks to capture the key feature information of the crew behavior, in order to reduce the influence of background noise and enhance the performance of the model. We analyzed the image features and found that the crew overboard behavior has different feature representations at different time steps. We expand the Channel Attention Mechanism (CAM) from 2D to 3D and introduce the time dimension, which constitutes the Channel Temporal Attention Mechanism (CTAM). The CTAM can effectively learn the key features at each time step, flexibly adjust the channel weights, and pay more attention to the crew behavioral characteristics at different time steps.

Figure 4 shows a sketch of the Channel-Time Attention Mechanism (CTAM). CTAM compresses the spatial dimensions of features by maximum pooling and average pooling, aggregating channel and time dimension features. The compressed features are learned by a multilayer perceptron to obtain two channel feature descriptions containing the time dimension. These two feature descriptions are then added and a weight vector is obtained by a sigmoid activation function. The application of the weight vector to each channel at every time step of the input feature map enables a weighted summation of channel features. This adjustment enables the model to fine-tune the importance of different channels at different time steps, which facilitates a more effective focus on features crucial to the current task. The introduction of CTAM reduces the interference of background information and improves the feature extraction capability of the model, which helps the model to focus more on the behavior of the crew.

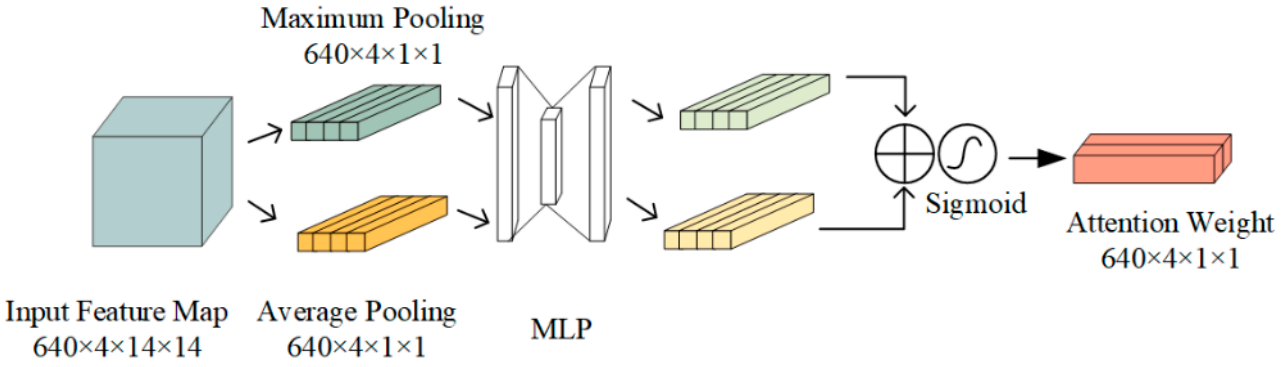


Figure 4. Channel-Time Attention Mechanism.

3.3. Activation Function

The Dropout layer in DR-Net randomly sets the output of some neurons to zero to reduce the risk of overfitting. Considering the influence of the activation function on these neurons, the model adopts the LeakyReLU activation function. This function does not output zero in the negative interval and retains the modified nonlinear characteristics, which can effectively solve the problem of neuron "death" in the ReLU activation function. In addition, the additional nonlinear characteristics introduced by the LeakyReLU activation function can effectively enhance the nonlinear modeling ability and network expression ability, and improve the robustness and generalization ability of the network.

4. Experiments

4.1. Self-Made Dataset

The performance of the algorithm is affected by the quality of the dataset in the task of detecting overboard behavior. To the best of the authors' knowledge, publicly available datasets of crew overboard behavior are very scarce. We collected surveillance videos related to drowning accidents through various methods, and obtained 10 video clips, including people falling into the water from boats and lakeside and riverside railings. We organized experimenters to simulate crew overboard behavior, and generated our own overboard behavior dataset based on accident surveillance videos and simulation experiment data.

In the simulation experiment, the height of the railings varies within the interval of [0.5, 1.5] meters. The positioning of cameras in the experiment is set at approximately 2 to 2.5 m from the ground, while the distance from the railing is varied within the range of [5, 10] meters. The falling behavior in the experiment includes both frontal and lateral falling in order to simulate the real crew overboard behavior. The simulation experiment obtains 1445 video clips, including 714 drowning behavior clips and 731 normal behavior clips, each lasting 2 to 7 s.

Figure 5 shows our dataset, which includes combinations of different railings (types and heights) and shooting positions (distance, height, angle). The dataset has a total of 1455 video clips, with 724 clips documenting overboard behavior and 731 clips documenting normal behavior. We analyze the behavioral characteristics in the dataset and conclude that (1) the information captured by the camera mainly consists of the back and side views of crew members, (2) personnel information is affected by various occlusions and there are few extractable features, and (3) there are some differences in the behavioral characteristics of members in different scenes.

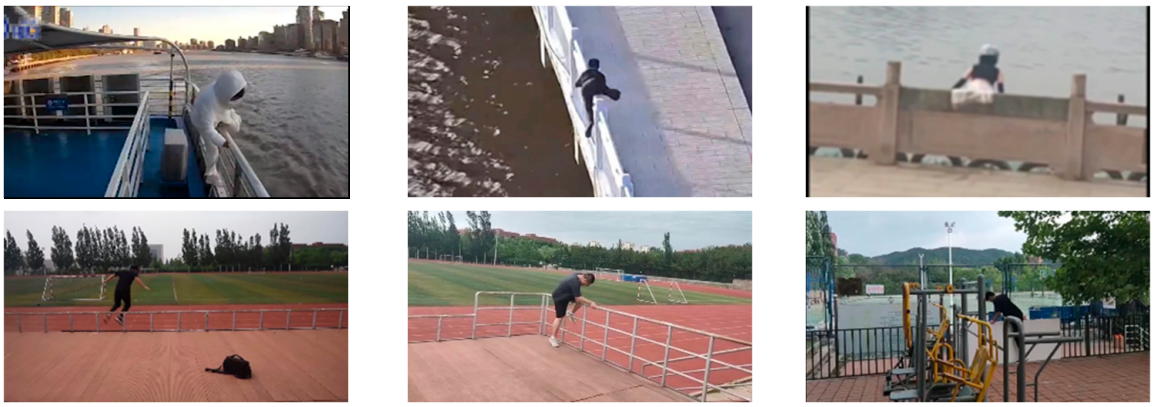


Figure 5. Illustration of our dataset.

4.2. Implementation Details

The experiments are conducted on a Windows 10 operating system, which utilizes an experimental platform equipped with an AMD Ryzen 9 5900HS CPU and an NVIDIA GeForce RTX 3060 GPU Laptop featuring 6 GB of video memory. The equipment described

is manufactured by ASUS in Shanghai, China. The modeling process is executed in a Python 3.8 environment with pytorch1.11.0. During the data preprocessing stage, each video frame is standardized to a resolution of 480 × 320. In the training phase, the input clips are randomly cropped into 32 × 224 × 224. The network parameters are optimized by standard SGD and the dropout layer is added with a 0.8 dropout rate. The initial learning rate is set as 0.01 and decreased by a factor of 10 every 50 epochs. The maximum number of epochs is set to 150.

4.3. Ablation Study

The DR-Net comprises two dense blocks and two residual blocks. In this section, we employ DenseNet-65 as the baseline, which consists of four dense blocks and incorporates pseudo-convolution internally. Five experimental groups are designed for ablation experiments to evaluate the feasibility and effectiveness of DR-Net structure, LeakyReLU activation function, and CTAM. The experiments are detailed in Table 1.

The experimental results show that, (1) compared with the baseline model, the accuracy of the DR-Net was improved by 6.3%. This is because the DR-Net employs dense blocks for extracting shallow features and residual blocks for extracting deep features, which reduces overfitting and enhances the model's feature extraction capabilities. (2) The DR-Net (LeakyReLU) model exhibits a 1% accuracy improvement over the DR-Net model using the ReLU activation function, which indicates that the LeakyReLU activation function enhances the model's nonlinear capabilities and expressive power. (3) The accuracy of the model improved by 0.3% and 0.4% with the addition of CTAM compared to the DR-Net and DR-Net (LeakyReLU) models without the attention mechanisms. The improved accuracy affirms that the CTAM enhances attention to critical channel information at each time step, strengthens the model's feature extraction capabilities, and improve its accuracy. (4) The DR-Net network with the LeakyReLU activation function and the introduction of the CTAM achieves the highest accuracy at 96.9%.

Figure 6 displays screenshots from two different normal videos. In Video 1, camera shake occurred during the recording, resulting in blurred image quality and partial loss of behavioral features. Additionally, some features of individuals are obscured by background equipment. In Video 2, the coverage area of background devices is larger, leading to the more extensive obscuration of individual features. For the detection of the two normal behaviors shown in Figure 6, DR-Net (LeakyReLU) + CTAM identifies them as normal activities (left), while DR-Net (LeakyReLU) detects them as falling behavior (right). This further underscores that the addition of CTAM is more effective in focusing on the behavior itself, reducing errors caused by background jitter and device obstruction.

The experiment results show that the model has a moderate increase in time–space cost, with a significant improvement in accuracy of 6.3%, and that the overall performance has been substantially enhanced. The addition of the LeakyReLU activation function and the CTAM results in very modest changes in the model's FLOPs and parameters, which has a negligible impact on hardware performance.

Table 1. Ablation experiment.

Methods	Dense Block	Residual Block	LeakyReLU	CTAM	Accuracy (%)	FLOPs (Mi)	Parameters (Mi)
Baseline	√	×	×	×	89.2	34,448.19	23.97
DR-Net	√	√	×	×	95.5	51,809.40	39.04
DR-Net (LeakyReLU)	√	√	√	×	96.5	51,809.40	39.07
DR-Net + CTAM	√	√	×	√	95.8	51,811.17	39.89
DR-Net (LeakyReLU) + CTAM	√	√	√	√	96.9	51,811.17	39.89

(**a**) Screenshot captured from video 1

(**b**) Screenshot captured from video 2

Figure 6. Screenshots of videos of normal behavior and detection results.

4.4. Main Results

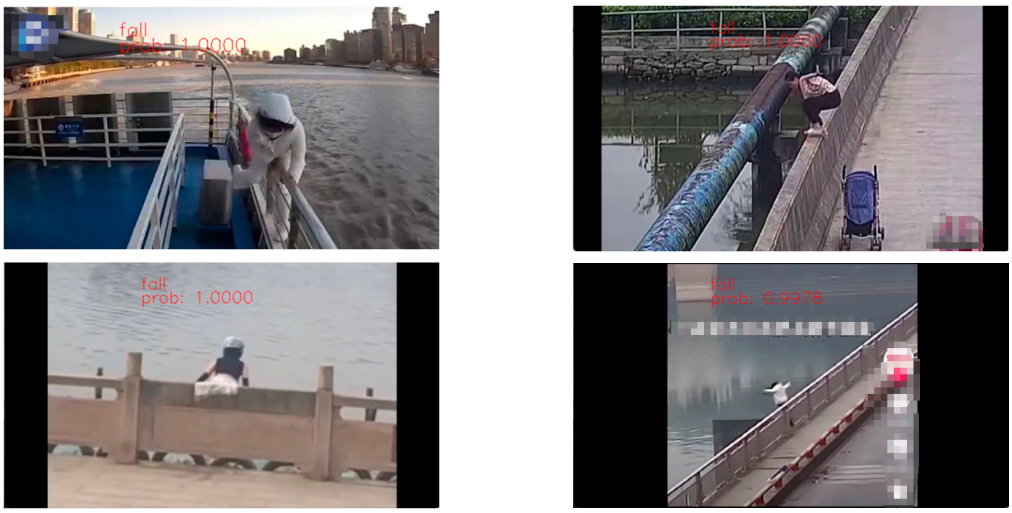
The method proposed in this paper differs from the concepts of the other man-overboard detection algorithms described in the introduction, and there are no relevant datasets available for comparative experiments. While the conventional algorithms described in our manuscript swiftly detect crew overboard situations and issue alarms, they often do so after the crew member has already fallen into the water. In contrast, CTDR-Net monitors crew behavior near the ship's outer railing, enabling the detection of tendencies or actions indicating a potential overboard situation before the crew members fall, which can enhance the timeliness of early warning. A comparison experiment was conducted between the method proposed in this paper and several mainstream action recognition models, which include P3D-63, ResNet-50, DenseNet-65, and R (2 + 1) d-50, to assess the performance of the proposed method. The detailed results, presented in Table 2, indicate the following:

1. Among the five action recognition network models, DenseNet-65 exhibits the lowest accuracy at 91.7%. Compared with this, the proposed method in this paper achieves a 5.2% improvement in accuracy, with the highest detection accuracy. The DR-Net network combines the advantages of dense blocks, residual blocks, and the CTAM in the feature extraction stage, resulting in a stronger feature extraction performance. The LeakyReLU activation function is used to enhance the generalization ability of the network to achieve preferable detection accuracy.

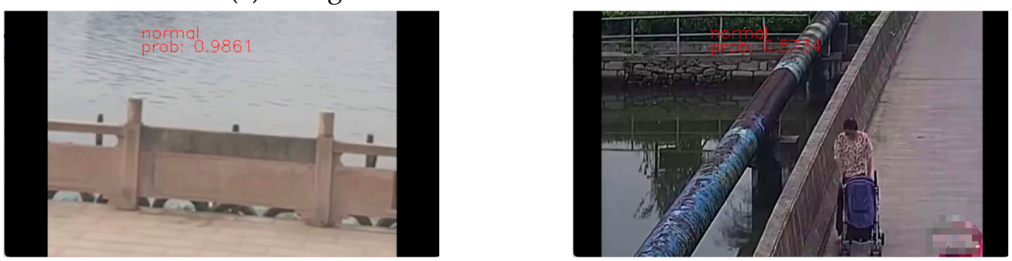
Table 2. Comparative experiment.

Methods	Accuracy (%)	FLOPs (Gi)	Parameters (Mi)	FPR (%)	FNR (%)
YOLO_WA [9]	86.9	-	-	-	-
Improve YOLOv4 [11]	98.68	-	-	-	-
P3D-63	94.1	17.37	20.53	7.75	4.11
ResNet-50	93.8	57.17	39.39	6.34	6.16
DenseNet-65	91.7	58.75	4.51	9.15	7.53
R (2 + 1) d-50	94.1	61.05	39.41	8.45	3.42
Ours	96.9	51.81	39.89	2.82	2.74

2. The parameters in the method proposed in this paper (39.89 Mi) are comparable to ResNet-50 (39.39 Mi) and R (2 + 1) d-50 (39.41 Mi), and higher than that of DenseNet-65 (4.51 Mi). The increase in the parameters is reasonable, considering the substantial improvement in accuracy. The FLOPs of the method in this paper (51.81 Gi) surpasses P3D-63 (17.37 Gi) but is lower than the other three models, which can achieve high accuracy in detecting crew overboard behavior with relatively low computational effort.
3. The method proposed in this paper demonstrates significantly a lower false positive rate (2.82%) and false negative rate (2.74%) compared to other algorithms. The detection results shown in Figure 7 further underscore the effectiveness of our algorithm in identifying overboard behavior. Figure 8 presents two instances illustrating false negatives and false positives in our algorithm. In the (a) video screenshot, individuals are heavily obscured and the algorithm fails to extract sufficient valid features, resulting in a false negative. Meanwhile, in the (b) video screenshot, the low image quality and extremely low pixel resolution cause the algorithm to extract incorrect features, leading to a false positive. These two examples show that, when the person is heavily occluded or the picture quality pixels are very low, this can result in extremely limited effective features that can be extracted. In such situations, the algorithms in this paper may not be able to perform effective detection.



(**a**) Recognition results of overboard behavior

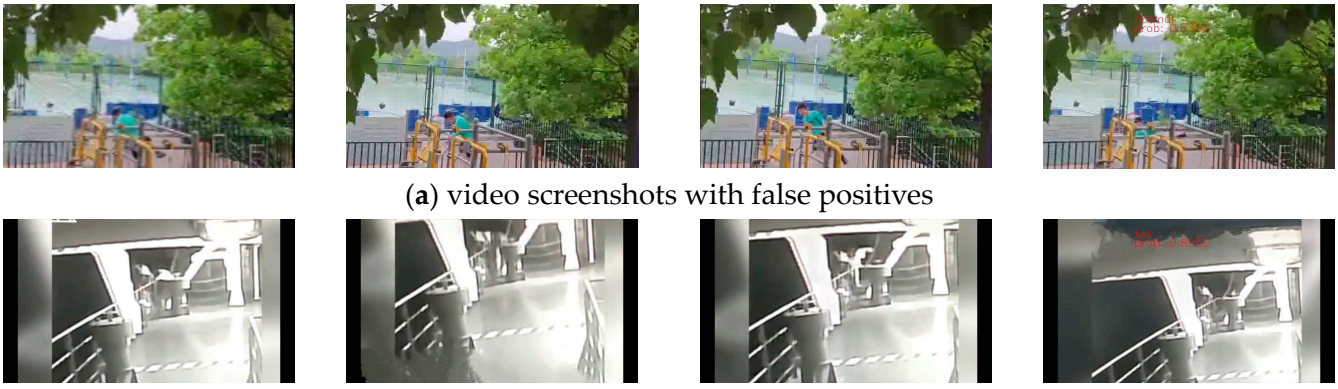


(**b**) Recognition results of normal behavior

Figure 7. CTDR-Net in video clip recognition results.

In order to evaluate the real-time performance of the proposed method, we conducted inference experiments on two distinct computers, and the specifications for each computer are provided in Table 3. The results of these experiments are summarized in Table 4. Notably, the average inference time on Computer 1 exceeded that of Computer 2, primarily due to the lower GPU performance on Computer 1. On the other hand, the average data loading time was observed to be lower on Computer 1. While the higher CPU performance on Computer 1 contributes to faster data loading, it is essential to consider that data loading time is a multifaceted metric influenced by various factors, including CPU, memory, and

storage card read/write speeds. The interaction of these factors affects the overall real-time performance of the algorithm. The average data loading time of the proposed method is only slightly higher than that of DenseNet-65, and the average inference time is slightly higher than other methods but lower than DenseNet-65. On Computer 1, the average total time for the proposed method is 162 ms, merely 2 ms more than P3D-63 and lower than other methods. On Computer 2, the average total time is 420 ms, comparable to P3D-63 (420 ms) and ResNet-50 (420 ms), and lower than alternative algorithms. These findings signify that the proposed method strikes a balance between accuracy and real-time performance.



(**a**) video screenshots with false positives

(**b**) video screenshots with false negatives

Figure 8. CTDR-Net false positives and false negatives in video screenshots.

Table 3. Computer configuration comparison.

Configuration	CPU	GPU	Operating System	VRAM	Python Version	PyTorch Version
Computer 1	AMD Ryzen 9 5900HS	NVIDIA GeForce RTX 3060 Laptop	Windows 10	6 GB	3.8	1.11.0
Computer 2	Intel Xeon(R) Bronze 3206R	NVIDIA GeForce RTX 3080 LHR	Ubuntu 22.04.2 LTS	12 GB	3.8	1.13.1

Table 4. Real-time performance comparison experiment.

	Computer 1			Computer 2		
Methods	**Average Total Time (ms)**	**Average Data Loading Time (ms)**	**Average Inference Time (ms)**	**Average Total Time (ms)**	**Average Data Loading Time (ms)**	**Average Inference Time (ms)**
P3D-63	160	110	50	420	387	33
ResNet-50	165	107	58	420	388	32
DenseNet-65	167	97	70	422	380	42
R (2 + 1) d-50	167	105	62	422	385	37
Ours	162	98	64	420	382	38

5. Conclusions

In this paper, we propose a Channel-Time Dense Residual Network (CTDR-Net) for detecting crew overboard behavior, including a Dense Residual Network (DR-Net) and a Channel-Time Attention Mechanism (CTAM). The Dense Residual Network (DR-Net) is proposed as the backbone network for feature extraction, which employs a convolutional splitting method to reduce the number of network parameters. The CTAM is used to enhance the expression ability of channel feature information, and can increase the accuracy of behavior detection more effectively. We used the LeakyReLU activation function to

improve the nonlinear modeling ability of the network, which can further enhance the network's generalization ability. The experimental results show that our method has an accuracy of 96.9%, striking a good balance between accuracy and real-time performance.

This work is a beneficial exploration in the field of crew overboard behavior detection. It is of great significance for improving emergency response capabilities for maritime disasters and ensuring the safety of crew life and property.

Author Contributions: Conceptualization, Z.L. and J.G.; data curation, K.M. and Z.W.; formal analysis, Z.L.; investigation, J.G.; methodology, Z.L.; project administration, L.D.; software, J.G.; supervision, K.M.; validation, Z.L., K.M. and L.D.; visualization, Z.W.; writing—original draft, J.G.; writing—review and editing, Z.L. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Qingdao Municipal Bureau of Science and Technology (grant number 22-3-3-hygg-3-hy).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu, C.Q.; Peng, J.B.; Jia, Y.G. Marine geohazards: Past, present, and future. *Eng. Geol.* **2023**, *323*, 107230. [CrossRef]
2. Sevin, A.; Bayilmis, C.; Ertürk, I.; Ekiz, H.; Karaca, A. Design and implementation of a man-overboard emergency discovery system based on wireless sensor networks. *Turk. J. Electr. Eng. Comput.* **2016**, *24*, 762–773. [CrossRef]
3. Örtlund, E.; Larsson, M. *Man Overboard Detecting Systems Based on Wireless Technology*; Chalmers Open Digital Repository: Gothenburg, Sweden, 2018.
4. Sheu, B.; Yang, T.; Yang, T.; Huang, C.; Chen, W. Real-time Alarm, Dynamic GPS Tracking, and Monitoring System for Man Overboard. *Sens. Mater.* **2020**, *32*, 197. [CrossRef]
5. Yan, L.; Pu, S.C.; Xu, F.; An, X.D. Study on the person water entry signal analysis and detection. In Proceedings of the 2021~2022 Academic Conference of Hydroacoustics Branch, Acoustical Society of China, Qingdao, China, 15 August 2022; Volume 4, pp. 425–428.
6. Pal, A.; Campagnaro, F.; Ashraf, K.; Rahman, M.R.; Ashok, A.; Guo, H.Z. Communication for Underwater Sensor Networks: A Comprehensive Summary. *ACM Trans. Sens. Netw.* **2022**, *19*, 1–44. [CrossRef]
7. Tsekenis, V.; Armeniakos, C.K.; Nikolaidis, V.; Bithas, P.S.; Kanatas, A.G. Machine Learning-Assisted Man Overboard Detection Using Radars. *Electronics* **2021**, *10*, 1345. [CrossRef]
8. Feng, D.W. Intelligent Identification and Positioning Rescue System for Falling into the Water by the Lake. Master's Thesis, Taiyuan University of Technology, Taiyuan, China, June 2022.
9. Wu, X.H.; He, Y.Z.; Zhou, H.; Cheng, L.; Ding, M.Y. Research on the personnel recognition in monitored water area based on improved YOLO v7 algorithm. *J. Electron. Meas. Instrum.* **2023**, *37*, 20–27.
10. Yang, Z. Research on the Detection Method of Pontoon Overboard Personnel. Master's Thesis, Jiangsu University of Science and Technology, Zhenjiang, China, July 2022.
11. Zhang, C.Y. Research on Key Technologies of Infrared Thermal Imaging Detection and Identification of People Falling into the Water at Sea. Master's Thesis, Dalian Maritime University, Dalian, China, June 2022.
12. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human Action Recognition from Various Data Modalities: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3200–3225. [CrossRef] [PubMed]
13. Luo, C.Y.; Cheng, S.Y.; Xu, H.; Li, P. Human behavior recognition model based on improved EfficientNet. *Procedia Comput. Sci.* **2022**, *199*, 369–376. [CrossRef]
14. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 8–13 December 2014.
15. Wang, L.M.; Xiong, Y.J.; Wang, Z.; Qiao, Y.; Lin, D.H.; Tang, X.O.; Gool, L.V. Temporal Segment Networks for Action Recognition in Videos. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2019**, *41*, 2740–2755. [CrossRef] [PubMed]
16. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
17. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

18. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
19. Qiu, Z.F.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
21. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018.