AquaScan: A Sonar-based Underwater Sensing System for Human Activity Monitoring

Haozheng Hou[1], Bowen Zheng[1], Sitong Cheng[2], Xiaoguang Zhao[2], Peiheng Wu[1], Lixing He[1], Yunqi Guo[2], Guoliang Xing[2*], and Zhenyu Yan[2*]
The Chinese University of Hong Kong, Hong Kong SAR, China
[1]{1155161507, bwzheng, wph19, 1155170464}@link.cuhk.edu.hk
[2]{sitongcheng, xgzhao, yunqiguo, glxing, zyyan}@cuhk.edu.hk

ABSTRACT

Human activity monitoring in the water is essential for pool management and drowning prevention. Existing camera-based solutions pose significant concerns about privacy and extra installation costs. Although sonars have been widely used for underwater sensing in open aquatic environments such as oceans and lakes, monitoring human activities with sonars in a pool setup is still challenging. In this work, we propose AquaScan, the first scanning sonar-based underwater sensing system for human activity monitoring. To overcome the low frame rate due to the sonar's physical limitation, we propose a novel scanning strategy and apply an image reconstruction method to accelerate the scanning speed without compromising the performance of motion detection. To overcome the dynamic interferences in the underwater scenario, we develop a novel signal processing pipeline based on a physical model to remove noises and localize human subjects. We further extract features like motion, time, and spatial information from sonar images and develop a state-transfer-based activity recognition system to recognize five common water activities, i.e., swimming, motionless, splashing, struggling, and drowning. We have deployed AquaScan on three public swimming pools for a total period of 94 hours. The evaluation results show that AquaScan can successfully recognize the five activities in the water with around 91.5%.

CCS CONCEPTS

• **Computer systems organization → Sensor networks**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

*Corresponding author.

ACM MOBICOM '25, November 4–8, 2025, Hong Kong, China

ACM ISBN 979-8-4007-1129-9/2025/11.
https://doi.org/10.1145/3680207.3723484

KEYWORDS

Underwater sensing system, Sonar, Human activity recognition

ACM Reference Format:
Haozheng Hou[1], Bowen Zheng[1], Sitong Cheng[2], Xiaoguang Zhao[2], Peiheng Wu[1], Lixing He[1], Yunqi Guo[2], Guoliang Xing[2*], and Zhenyu Yan[2*]. 2025. AquaScan: A Sonar-based Underwater Sensing System for Human Activity Monitoring. In *The 31st Annual International Conference on Mobile Computing and Networking (ACM MOBICOM '25), November 4–8, 2025, Hong Kong, China.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3680207.3723484

1 INTRODUCTION

Continuous monitoring of human activities in swimming pools is vital for effective pool management, training purposes, and ensuring safety. A report shows that almost one-third of drowning deaths in the United States occur at lifeguarded pools [51], making pool management critical. Lifeguards may experience attention lapses and have an obstructed line of sight (LoS) due to their positioning above the water, making it challenging to distinguish between normal behavior and potential drowning incidents.

Moreover, existing solutions fail to address the need for continuous, long-term individual monitoring in many aquatic environments. Some pool management systems install cameras at the bottom of the pool or above the pools [16, 28, 38, 55, 60, 64], but they are prone to being compromised by varying lighting conditions and raise significant privacy concerns from the public. The camera-based system requires installation at the pool's bottom and the edges, raising significant concerns about privacy from the public. Moreover, it suffers from dynamic lighting conditions and foggy water. Additionally, motion recognition solutions for swimmers that rely on wearable devices [22, 27, 33, 44, 47, 53] are often obstructive, intrusive, and lack scalability. In summary, current solutions fail to enable the continuous monitoring of human activities in water without incurring high deployment costs or invading privacy.

Sonar technology is widely used for underwater sensing, from deep ocean monitoring [13, 23, 36] to robust underwater

communication. The ability of acoustic waves to propagate through water and capture the movement of subjects including human bodies makes sonar a promising sensor modality for underwater sensing. However, most conventional sonar solutions are designed for long-range sensing, which can be prohibitively expensive and potentially unsafe in swimming pool environments. Recent research has explored the use of smartphone speakers and microphones for underwater sensing [6]. Nevertheless, such devices do not possess the capability to extract the fine-grained features necessary for precise activity recognition.

In this work, we develop the first activity monitoring system with a new generation of sonar, called *scanning sonar*. Scanning sonar features a motor-driven transducer that can pivot to specific angles, allowing it to emit sound waves in a narrow beam, typically spanning 2.22 gradians (grads) horizontally and 27.78 grads vertically. With the ability to rotate the transducer through full scanning, the motor enables comprehensive scanning coverage of the swimming pool area. The output of scanning sonar is a low-resolution 2D sonar image showing the horizontal space, which also costs significantly less than the other high-end imaging sonars.

These characteristics make scanning sonar an ideal solution for underwater navigation and mapping [17]. However, sensing human activities with scanning sonars faces three major challenges: First, the sonar's frame rate is limited by the physical rotation speed of the motor and acoustic echo return time, resulting in potential delays in detecting rapid activities. Second, the sonar images contain significant dynamic noises from various sources, such as water surface reflections, movements from other swimmers, and environmental factors like wind or rain. Third, activity recognition with sparse information is inherently challenging due to the limited information available for feature extraction. Common activities in the swimming pool include three safe activities: moving, standing, splashing, and two dangerous activities: struggling and drowning. These five activities are important for pool management and danger alarming, which is hard to recognize directly from sonar images since the 2D nature of sonar images compresses all vertical information onto a horizontal plane, complicating the extraction of 3D motion features and differentiation between similar activities.

To address the challenges outlined above, this paper introduces AquaScan, a novel sonar-based underwater sensing system designed for non-intrusive monitoring of human activity in swimming pools. Our system utilizes acoustic waves to accurately detect and classify various human movements, providing an effective solution that also respects the individual's privacy. AquaScan's approach to human activity monitoring is multi-faceted, incorporating advanced signal processing techniques, state-of-the-art machine learning algorithms, and an innovative scanning strategy that collectively overcomes the limitations of traditional sonar systems in a pool environment. By integrating these elements, AquaScan delivers a system capable of high accuracy and real-time monitoring, offering a significant advancement in the domain of aquatic safety and pool management. The main contributions of this paper are as follows:

- We propose an innovative intermittent scanning strategy, which is the first to effectively balance both the frame rate and detection performance of the underwater activity recognition system while also minimizing false detection and miss detection in object detection.
- We develop a physical-aware adaptive noise removal algorithm that effectively filters out environmental and operational interferences, ensuring the clarity and reliability of sonar images.
- We analyze the common human activities in the swimming pool and introduce a multi-dimensional feature extraction and state-transition framework for activity recognition, capable of distinguishing between various movements and identifying potential accidents or hazardous situations.
- We implement and evaluate AquaScan in three public swimming pools, collecting 94 hours of data[1]. The evaluation results show an average 91.5% accuracy for activity monitoring in 9.18 seconds. Our scanning strategy increases the frame rate by 1.83 times. Compared with the existing methods, we achieved 42.4% higher accuracy.

The structure of the paper is as follows: Section 2 reviews existing work. Section 3 measures scanning sonar's capabilities and challenges. Sections 4 and 5 describe the design and evaluation of AquaScan, respectively. Finally, Section 6 discusses the wider impact, limitations, and future research.

2 RELATED WORK

2.1 Monitoring Human Activities in Water

Various sensing modalities, such as vision, Inertial Measurement Units (IMUs), and acoustics, have been employed in aquatic sensing applications.

Cameras have been used for swimming style detection by capturing the swimmers' motion [38, 55] and gesture recognition for underwater human-robot interaction [60], but they only work for short-range recognition under sufficient luminance. Despite swimming detection, cameras can be used to detect drowning through object detection and skeleton recognition [16, 28, 64], which is a more challenging task due to the complexity and heterogeneity of drowning.

Wearable devices equipped with various sensors have been utilized for detecting underwater activities. IMUs have

[1]The data collection and experiments have been approved by the authors' institutional review board.

been integrated into smartwatches [53], body-mounted devices [33, 44], and headgear [27] to identify swimming styles. Other sensors, like radio [22], heart-rate tracker [47], or integrated multi-sensor system [35] are also proposed to detect drowning by some pre-defined feature. However, they failed to accurately distinguish drowning from other water activities like splashing or diving due to their inability to capture full-body motion. Moreover, it is not practical to ask all swimmers are accessible through wearable devices.

Compared to the above modalities, acoustics represents the promising modality for underwater detection due to relatively low attenuation underwater, which will be discussed extensively below. [59] shows several underwater machine-learning applications based on acoustic data, including object detection, seafloor detection, and target classification. Specifically, previous works utilize Doppler-frequency-shift [24, 34], convolutional neural network [34] to recognize drowning but ignore the water and multi-user interferences. [21] demonstrates sonar-based drowning detection, which is the closest work to our system. However, it provides limited analysis of human activity patterns, thereby limiting the target to two basic classes that can't cover sufficient cases. Aquahelper [61, 62] build an SOS transmission system by smartwatch; although promising, it is impractical to rely on potential drowning victims to activate SOS signals.

2.2 Sonar and Acoustic Sensing

SOund NAvigation and Ranging (Sonar) has been extensively explored for underwater applications, which uses acoustic sensing to detect underwater objects with relatively low attenuation. It has three types: side-scan, multi-beam imaging, and rotary scanning sonar.

Side-scan sonars, known for their narrow and tall beams, excel at capturing detailed images over long distances, making them effective for geological and structural mapping [5, 7, 30]. However, their fixed view and narrow beam width limit their coverage area, potentially missing nearby swimmers in a pool setting.

Multi-beam sonars mitigate this by emitting multiple beams to scan and image confined areas, commonly used to track underwater objects and marine life [13, 23, 36]. Yet, their constrained scan range can lead to blind spots, necessitating additional sensors and thus escalating costs.

Rotary scanning sonars, which employ a motor-driven single-beam transducer to achieve 360-degree coverage, are predominantly utilized for undersea environmental monitoring [18, 42, 43, 46]. Wide coverage and lower cost than multi-beam sonars make rotary scanning sonar suitable for human activity sensing.

Although there are various applications of acoustic sensing, including fall detection by Doppler shift [37], limbs and torso detection [4], 3D pose reconstruction with RGB images [63], localization [39, 52], finger gesture [50] and underwater localization [6], all these above techniques utilize frequency-domain or phase-domain features for activity detection, requiring access to raw data. Unfortunately, the commercial sonar doesn't support it, which motivates us to design our new system for human activity monitoring. [23, 26] shows the potential of marine animal detection with acoustic sensing. However, these works cannot be applied in our system due to the limited sensing range and lack of design for multi-subject detection.

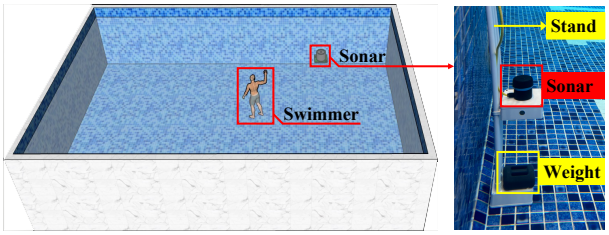


Figure 1: Measurement setup in a public pool.

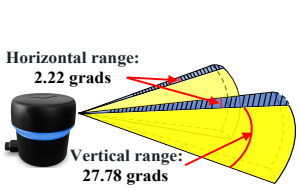


Figure 2: Working principles of the scanning sonar.

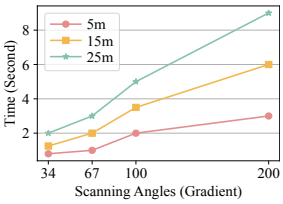


Figure 3: Scan speed with different sonar settings.

3 A MEASUREMENT STUDY OF SCANNING SONAR

While sonars are widely used for sensing in open aquatic environments, their adoption for accurate human activity monitoring in compact water spaces like swimming pools is still emerging. Scanning sonar equips a motor to rotate the sonar transducer for a larger coverage. This section measures its performance in a public swimming pool and analyzes its capability of distinguishing human activities in the water.

3.1 Scanning Sonar in Underwater Sensing

To assess the performance of the scanning sonar underwater, we attached a sonar to the edge of a public pool.

Fig. 1 demonstrates our deployment setup. The sonar, deployed at 1.5 meters underwater, is connected to a Raspberry Pi using an ethernet cable to collect data.

Working Principles of Scanning Sonars. Ping360 Scanning Imaging Sonar produced by BlueRobotics [1], is chosen for our measurement study. The sonar uses 5 watts of power as the Maximum Power Consumption. Fig. 2 shows the working principle of the scanning sonar. In each *scan*, the sonar

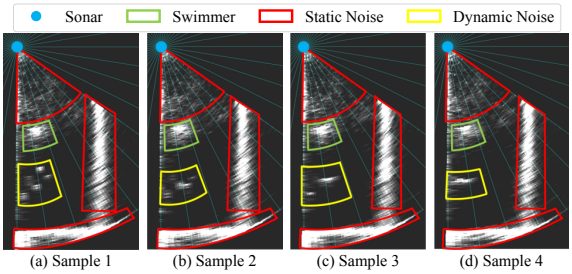


Figure 4: Four example images from a scanning sonar.

transmits a 750 kHz acoustic wave to measure distances based on the time it takes for the signal to return. It scans a sector with a horizontal angle of 2.22 grads and a vertical angle of 27.78 grads, rotating in different directions to fully scan all consecutive sectors. The output is a 2D greyscale image where each pixel indicates the intensity of reflections.

Sonar Scanning Time. We measure the rotation speed of the scanning sonar to estimate the delay in scanning the whole pool with sensing ranges of 5 m, 15 m, and 25 m and scanning 34, 67, 100, and 200 grads. As shown in Fig. 3, larger scanning areas and ranges increase the scanning time. For example, it takes the sonar 9 s to scan a pool with a 25 m sensing range and 200 grads (180 degrees) scanning area, which is 0.111 frames per second (FPS). This delay occurs as the sonar waits for acoustic echoes which causes poor tracking performance and low recognition accuracy due to capturing fewer frames per movement.

Static and Dynamic Noise in Sonar Images. Fig. 4 presents four sonar images of a volunteer standing in a pool, each showing a 60 grads scan. The blue dot marks the position of the sonar, while the green box indicates the location of the subject. Red boxes highlight static noises caused by the edges and bottom of the pool. Yellow boxes represent random dynamic noises from sources such as water surface waves and reflections between objects and pool edges. Observations suggest that pre-scanning the background to eliminate static noises is beneficial. Some dynamic noises are less energetic and may overlap with other human subjects. Given the unpredictable nature of water movements and reflections, it is desirable to design an adaptive noise removal method to ensure accurate detection of human subjects.

3.2 Human Activities in the Water

Human aquatic activities can be categorized as either voluntary or instinctive. Voluntary actions, such as intentional swimming and purposeful splashing, are consciously initiated by individuals. On the other hand, instinctive actions, like struggling to remain buoyant or losing control in dangerous situations like drowning, occur involuntarily and surpass conscious control [56].

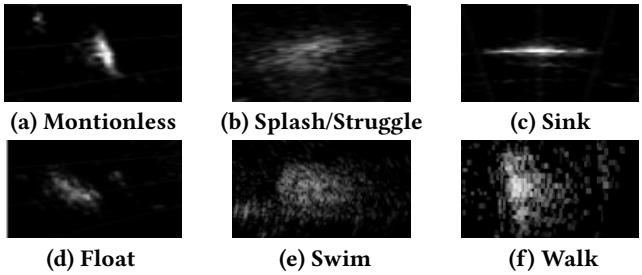


Figure 5: Sonar images of six human activities.

We collected the scanning sonar data when a volunteer engaged in various aquatic activities[2]. Fig. 5 presents six sonar images captured when a volunteer is motionless, struggling, sinking, floating, swimming, and walking, respectively. The activities with minimal motion in Fig. 5a, 5c, and 5d generate more concentrated clusters. In contrast, activities such as those in Fig. 5b result in larger clusters due to increased movements. Fig. 5b, 5e and 5f further show that when the arms are open or waving, the image shows more scatters around the object. Even though sonar images may not clearly show poses or complex movements, we observe that the distribution of echoes reflects the subject's motion. The body parts demonstrate higher energy in the sonar images, while the arm movements create scatters near the body signal. Splashing or struggling can be classified through dispersed echoes due to motion. Sinking, being motionless, and floating share similar features that classify them as motionless states. Given that swimming and walking lack distinctive features within one frame, obtaining the location of swimmers becomes crucial for assessing subjects' movements.

We collected data on three common activities: swimming, patting, and standing, as well as on two hazardous actions: struggling and drowning. We trained two deep neural networks (DNNs), Resnet18 and video activity analysis [10], using our self-collected sonar image dataset. Each input sample consists of 3 consecutive frames of sonar images so that the models can capture the temporal information. Both these two models achieve low accuracy at 0.397 and 0.25.

This measurement study highlights three pivotal challenges faced by our sonar-based system: scanning speed, dynamic noise, and the recognition of complex swimming activities. Firstly, the current sonar system requires 6 seconds to scan a 200-grad area with a 15-meter range, a pace insufficient for real-time application, leading to delayed activity recognition. Secondly, the presence of dynamic noise significantly increases false positives, thereby reducing reliability and increasing the risk of false alarms. Thirdly, the diverse aquatic activities of individuals complicate accurate identification and categorization, necessitating enhanced recognition capabilities.

[2]All experiments have been approved by the IRB of the author's institute.

4 AQUASCAN DESIGN

4.1 System Overview

To overcome the aforementioned challenges, AquaScan is designed to support continuous, timely, and accurate monitoring of human aquatic activities by utilizing scanning sonars with high scanning frame rates.

Fig. 6 shows the overview of AquaScan's design. First, AquaScan intermittently controls sonar scanning for low-latency scanning. Second, AquaScan eliminates static noise and reconstructs sonar images to compensate for intermittent scanning. Then, to detect and locate human subjects in sonar images, AquaScan utilizes a dual-branch noise removal pipeline with physical information. After merging localization and detection results, AquaScan recognizes activities using a multidimensional state machine by extracting time, motion, and spatial features. The recognition includes moving, motionless, possible drowning, struggling, and splashing.

4.2 Intermittent Scanning Strategy and Image Reconstruction

Our measurements in Section 3.1 show that a full scan with 15 m range takes 6 s. This duration could potentially lead to missing the detection of critical activities. To accelerate the sonar scanning, we propose a scanning strategy of skipping scanning angles intermittently to improve the data framerate. The idea is that the sonar scans the first x consecutive grads every y grads for each image. In the rest of the paper, we use *x/y* to represent *scanning first x gradians in y gradians*.

Since the intermittent scanning strategy generates lines of empty pixels in sonar images, we design an image reconstruction method for interpolation.

Existing popular image processing methods often use deep learning (DL) (e.g., Variational Auto-encoder (VAE) [32], Unet [49], masked autoencoder (MAE) [19]) for image generation. However, these kinds of DL work well due to much redundant information in RGB images. MAE works well on images with random sampling, but intermittent scanning produces block-mask images where MAE degrades performance [19]. Collecting the ground truth is hard for model training, which is also a large overhead. So, we employ an efficient interpolation algorithm that recovers skipped signals by calculating the average of the nearest existing pixels.

To motivate our choice, we collected full scan images and downsampled them to simulate intermittent scanning. We trained an Unet-shaped model [49] for sonar image reconstruction due to its great capability of capturing image features and calculated the mean square error (MSE) between full scan images and reconstructed images. The normalized MSE of our interpolation method is 0.00097, which is lower than the MSE of Unet (0.001). Considering that training a DL model requires large amounts of data, our interpolation method can achieve better performance without any data collection overhead.AquaScan also scans the background during the spare time to remove static noises, such as the pool's edges, bottom, and lane lines.

The intermittent scanning strategy brings benefits to both object detection and recognition in three folds: (1) Skipped angles can reduce each frame's scanning time to improve the sampling rate, which benefits detecting motion and motionless activities. (2) Faster scanning provides more sampling points on each swimmer's trajectory for better tracking performance. (3) The intermittent scanning strategy can skip slender noise to reduce false detections.

4.3 Dynamic Noise Removal and Object Detection

In the measurements described in Section 3.1, dynamic scattered noise has been observed in the sonar images, varying over time. To mitigate this issue, we implement a median filter [25], an effective noise-reduction method to remove high-frequency noise and extract edges, which works by replacing the central pixel in a kernel with the median value of the surrounding pixels. A crucial aspect of median filtering is the choice of kernel size that determines the number of pixels for calculating the median value used in averaging each pixel. To further understand the effect of varying the median filter's kernel size, we apply the median filter on our dataset used in Section 3.2. The results in Fig. 7 show that an increase in kernel size correlates with a reduced false detection rate but can cause a higher miss rate due to the stronger denoising effect, which may inadvertently remove important features of subjects. A smaller kernel size, however, preserves the clusters of subjects more effectively but also retains a greater amount of noise. This presents a significant challenge for subsequent object detection tasks. Especially when the kernel size is larger than 13, the miss rate increases dramatically. This is because kernels over 13 smooth the subject with extra surrounding information and decrease the size. Hence, it is desirable to adopt two kernel sizes to achieve both dynamic noise removal and object detection.

Based on our experiments with various kernel sizes using the median filter, we developed a novel dual-branch method for both noise removal and object detection. The processing pipeline of this method is illustrated in Fig. 8. Our approach initially focuses on eliminating weak echoes that are likely to be noises and overlap with the clusters representing swimmers. Subsequently, we employ median filtering with two distinct kernel sizes: one tailored for dynamic noise removal and the other for object detection. We use a search algorithm to determine the kernel sizes at runtime. The final step integrates the results from both branches, yielding sonar images annotated with bounding boxes that identify human subjects. It is important to note that our methodology incorporates

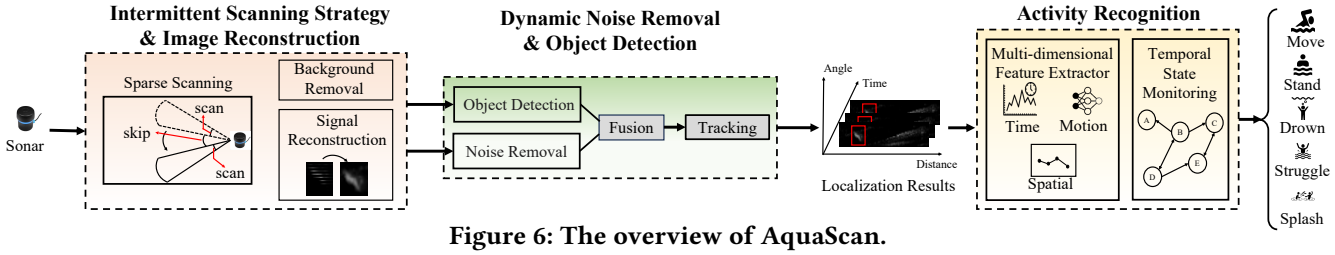


Figure 6: The overview of AquaScan.

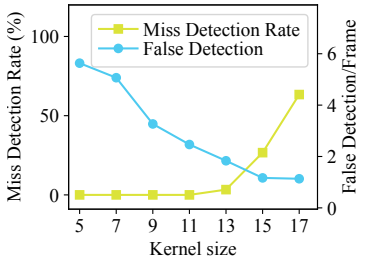


Figure 7: Performance of Median Blur with variable kernels sizes.

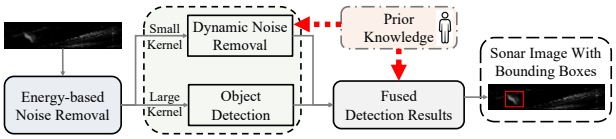


Figure 8: The processing pipeline of the dual-branch noise removal and objective detection.

prior knowledge about typical human dimensions in both the dynamic noise removal and the results fusion process.

For noise removal and object detection, we implement a pre-process for removing the scattered echoes that mix nearby subjects or generate fake subjects. We set different thresholds for different distances due to the attenuation of the acoustic. Note that different thresholds should be set accordingly for tests in new swimming pools and deployment positions. We then upscale the sonar images from a resolution of 400×500 to 1200×1500, allowing our density-based clustering algorithms to more effectively differentiate between clusters. It is worth noting that image resizing is beneficial for object detection only when the noise has been adequately reduced. Excessive noise can cause nearby clusters to merge, resulting in a higher false positive rate.

4.3.1 Physical-aware Dynamic Noise Removal. To optimize parameters for the median filter, we propose a search strategy leveraging prior knowledge of the human body size. This knowledge imposes a significant constraint on generated bounding boxes, as noise typically exhibits abnormal shapes and sizes, such as occupying too large or too small areas. We design a physically-aware adaptive tuning algorithm as shown in Algorithm 1. By incorporating physical knowledge, we can tailor the median filter to be adaptive to each image.

Our adaptive pipeline is designed to preserve the clusters of human subjects. Thus, we start with a small kernel size, recorded as K_{init}. If the generated bounding boxes exceed the body size threshold, it indicates that our algorithm's noise reduction is insufficient, leading to subjects merging with noise. As the kernel size increases, improved noise reduction capabilities facilitate the removal of noise and clearer delineation of subject cluster boundaries. Kernel size incrementation stops once all bounding boxes conform to the physical constraints, with a maximum kernel size limit set at K_{max} to prevent excessive loss of information and high latency. The threshold (T_e) represents the estimated maximum size of a human body. Assuming a person is lying flat on the water's surface, the area is calculated as height times width. Given an assumed height of 2.72 m [58] and a waistline of 3.02 m [15], the width, considering the human body as an approximate cylinder, is about 0.96 m. However, in real-world measurements, humans do not occupy such large areas, so this figure serves as the upper limit for the bounding box size. We convert this maximum area into a fixed bounding box size threshold, which is calculated when the human stands close to the sonar, making taken-up areas large. Theoretically, the bounding box size for a human object decreases with increasing distance from the sonar, and large T_e can prevent excessive denoising on human objects at long distances, making our theoretical maximum threshold applicable and reasonable. We also estimate the minimum size of bounding boxes by measuring them at different distances, finding that all swimmer clusters span more than 10 gradians and exceed the length threshold. Bounding boxes failing to meet these angle and width thresholds are considered noise. All these parameters should be tuned according to the small-scale real-world experiments in the pools.

4.3.2 Cluster-based object localization and bounding box generation. We apply a clustering algorithm DBSCAN [11] to localize the targets on the image, which groups together data points that are closely packed while effectively identifying outliers as noise. We utilize real-world parameters like the diameter of humans' occupied area as the reference to determine the parameter of DBSCAN, which can be fine-tuned during the deployment. After obtaining the bounding box for the cluster, we calculate the amplitude-weighted average coordinates as the location of the body. Next, the filtered

Algorithm 1 Search for the optimal kernel size

Define $f(x)$ as the clustering method, where x is the image.
Define $G(x, k)$ as the median blur function, where k is the kernel size.
Define $S(a)$ as the function to calculate the size of a bounding box a.
Initialize kernel size $k = K_{init}$, threshold T_e, where T_e is the maximum size of a bounding box representing a human.
while $k \leq K_{max}$ **do**
  Apply median blur $x_{temp} \leftarrow G(x, k)$
  Cluster objects $Obj \leftarrow f(x_{temp})$
  if any $S(Obj_i) > T_e$ for $Obj_i \in Obj$ **then**
    $k \leftarrow k + 2$ {Increase kernel size}
  else
    return k {Optimal kernel size found}
  end if
end while
return k {Return the largest considered kernel size if no smaller optimal size is found}

bounding boxes generated through the dynamic noise removal part (denoted as (a)) and object detection part (denoted as (b)) will be merged by iteratively comparing the ratio of overlapped areas on each corresponding bounding box from (a) and (b). If no bounding boxes can be matched, we will choose to keep the bounding boxes generated from (a) since keeping all the potential swimmers is more important than removing noise. One concern is if (b) misses objects, bounding boxes merging cannot be performed. In nearly all cases, if swimmers are missed, noise and other rest correct clusters are removed, or there is a lower IoU with bounding boxes from (a) due to excessive denoising ability.

4.3.3 Dynamic coverage with object tracking. AquaScan tracks each subject by predicting the potential locations of each trajectory. Distinguishing the specific subject from other nearby subjects is hard. To maintain the trajectory's correctness when mismatching the wrong subjects, we apply Multiple Hypothesis Tracking (MHT) [31] and process each trajectory and subject based on the number of nearby subjects and trajectories. First, we will split the trajectory trees into related trees and non-related trees according to whether there are nearby subjects. Similarly, subjects are split into related subjects that can be matched with nearby traces and non-related subjects that will be seen as new traces due to failing to match related traces. Next, we will match trajectories with unique matched subjects. For the remaining subjects and trajectories, we propose a scheme based on N_{trace} and $N_{subjects}$ (N means the number of trajectories or subjects). When $N_{trace} \geq N_{subjects}$, we assume that each subject can be uniquely assigned to a trajectory. Therefore, we first evaluate the distance between the subjects' locations and the predicted trajectories' locations. When $N_{trace} < N_{subjects}$, we allow the nearby subjects to share one trajectory.

Table 1: Activity definition.

Activity	Description
Moving	Clear change in location
Motionless	Minimal change in location, low-intensity motion
Struggling	Continuous high-intensity motion with minimal location change
Splashing	Short-term high-intensity motion with minimal location change
Drowning	A state after long-term struggling or extremely long-term low-intensity motion with minimal change in location

4.4 Multi-dimensional Activity Recognition

In our system, we aim to detect three **safe activities**: **motionless**, **moving**, and **splashing**, as well as two **dangerous activities**: **struggling** and **drowning**. The selection and definition of these activities are extracted from the guideline in [8]. The features distinguishing each activity are detailed in Table 1. Moving shows a clear change in location. The activities of minimal movement and motion are categorized together under the motionless class. Splashing is characterized by vigorous motion in the water while still maintaining control for a limited duration, unlike struggling, which also involves intense motion but suggests a loss of control. Drowning is typically a state that ensues struggling or long-term motionlessness, such as drowning caused by drunkenness.

Challenges. The human aquatic activities are highly diverse, which degrades the performance of the end-to-end DL model. For example, splashing and drowning are similar in a short time but have different developments in a long time. Sonars deployed at different pools are set at different scanning ranges and distances, which can lead to unpredicted frame rates. To solve these two challenges, we propose a three-dimensional feature extractor to extract the stable spatial, temporal, and motion features for each subject. A finite-state machine is designed for feature smoothing and long-time activity inference.

To enhance the recognition of daily activities and potentially dangerous situations, we integrate features from three domains: motion, temporal, and spatial. Our approach diverges from other activity recognition tasks that typically employ complex deep learning models. Instead, we utilize low-dimensional signals capturing motion, temporal, and spatial features that allow us to construct a definitive state-transition graph to deduce the activities. This methodology enables our system to monitor swimmers effectively, offering

superior generalization capabilities and more transparent physical interpretability.

We note that the definition of drowning in this paper serves as a reference, which can be customized according to various factors such as environment, pool settings, and users' requirements. AquaScan is designed to capture essential aquatic human features for recognizing a range of activities, accommodating a broad spectrum of use cases.

4.4.1 Motion features. Our measurements show that one human subject consists of echoes from the body and limbs. Their temporal change of shapes and sizes is indicative of the motion status, which is not affected by human diversity.

We use a ResNet-18 [20] model to discern these features and determine if a person is engaged in vigorous activity by analyzing three consecutive frames. We train the model with full-scan images and augmented images generated by simulating intermittent scanning to make the model capture the sonar global and local features better. This approach strikes a balance between information sufficiency and sensing delay. The model predicts whether the objects are in still or motion.

4.4.2 Spatial features. To categorize activities as moving, motionless, or splashing, we examine patterns of movement. The first step is to distinguish between moving and motionless states, which is usually achieved by tracking objects and measuring their location changes. However, struggling and splashing can also cause minor shifts in location, which may be mistaken for movement when only considering overall changes. Therefore, a more precise analysis of movement is needed. Inspired by K-means [41] for improved spatial feature extraction, we track an object's multiple locations in past sonar frames. We compute the centroid of a swimmer's positions over a time window T_{window} which contains no more than 6 consecutive samples and the mean distance d from the centroid to the swimmer's locations. A larger d indicates significant location variation within T_{window}. We evaluate the distance D with the last location, the mean distance d, the ratio R of D to d indicating movement direction, and the IoU of consecutive bounding boxes. Thresholds D_{max}, D_{min}, d_{max}, d_{min}, R_{min}, and IoU_{max} are set to categorize movement. Objects with an IoU or the ratio of the overlapped area on the smaller one above IoU_{max} or D and d below D_{min} and d_{min}, respectively, are deemed stationary. Conversely, a d greater than d_{max} signifies apparent movement. Swimmers are marked as moving if D exceeds D_{max} and R surpasses R_{min}, indicating a definitive movement trend. Other cases will be marked as stationary. We smooth the moving state in the sliding windows in one trajectory.

4.4.3 Time-domain features extraction. To refine the distinction between splashing, struggling, and drowning, we incorporate time-domain features. We begin by arranging the motion states in chronological order, then calculate the average duration of continuous motion, denoted as $\bar{F}_{motion}$. As previously established, individuals in danger often exhibit uncontrolled and persistent attempts to stay afloat. Therefore, a lower $\bar{F}_{motion}$ suggests better body control and a reduced likelihood of struggling. Conversely, a higher $\bar{F}_{motion}$ indicates a greater probability of struggle. To calculate $\bar{F}_{motion}$, we maintain a list of motion detection results within a time window that is suggested not to exceed 30 seconds and can calculate the ratio of current motion in this window. Drowning is typically a sequential state of struggling. Thus, when the confidence for motionlessness surpasses that for motion following a struggle or the struggling is maintained above a threshold $T_{struggling}$, we transition to a state of drowning. Informed by lifeguard insights, children's drownings can be deceptively tranquil; hence, we implement a rule-based link between prolonged motionlessness and drowning. If motionlessness persists for an exceedingly lengthy period $T_{motionless}$, it is also classified as drowning.

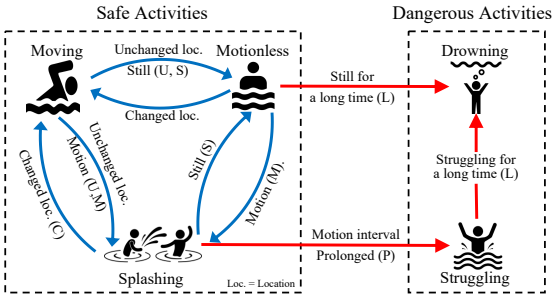


Figure 9: Multi-dimensional finite-state machine for activity monitoring.

4.4.4 Finite-state machine for activity monitoring. This section details the process of monitoring states and outlines the feature extraction for motion, temporal, and spatial aspects. The starting point will only be given splashing and motionless due to lack of spatial features. The collected images are processed through a multi-dimensional feature extractor, yielding four indicators: (1) Location changed (C) or unchanged (U), (2) In motion (M) or still (S), and (3) maintain one activity for a long time (L) or prolonged interval of motion (P). These indicators inform the transitions within the finite-state machine depicted in Fig. 9, enabling the progression from one state to another. To mitigate the impact of performance fluctuations on continuous activity monitoring, we apply a sliding window-based majority voting with no more than 5 samples on one trajectory. The first two samples will be smoothed with limited past state information so we apply majority voting to them again only when their following two states change. Utilizing a voting approach is justified by our objective to track activities over time, where sustained states provide significant insights.

5 EVALUATION

We have extensively evaluated AquaScan in real-world environments. First, we present the implementation details, evaluation setup, metrics, and baselines. Then, we describe the hyperparameter settings for AquaScan. Following that, we display the results of an end-to-end evaluation experiment. Subsequently, we assess the performance of the scanning strategy and object detection. Finally, we demonstrate AquaScan's performance under various impact factors.

5.1 Implementation and Experiment Setup

Hardware. We deploy AquaScan system with one or multiple Ping360 sonar units [2] introduced in Section 3.1. Fig. 10 shows two example deployments. Each sonar unit is connected to a processing unit above water via a Gigabit Ethernet cable. This processing unit includes a Raspberry Pi 4B with 8GB RAM, which streams the data to our server using Wi-Fi for storage and further analysis. The server runs on an AMD Ryzen 9 7950x3D CPU and an Nvidia RTX 4090 GPU.
Software. We implement the AquaScan software in Python for data processing and sonar control. Our code leverages Numpy and OpenCV for preprocessing sonar data and object detection, respectively. The recognition model of pool activities is trained using PyTorch library 1.13.0(version number) [45]. ResNet18 whose size is 44.8MB is trained for motion detection. This compactness ensures the practicality of deploying our solution in real-world scenarios. We have open-sourced our codes at [3].
Setup of Data Collection. We deployed one or two sonars in three public pools: a 25 m × 9 m pool (Pool A) designated for training and validation, and two larger pools, 30 m × 15 m (Pool B) and 50 m × 25 m (Pool C), used for testing. In Pools A and B, the sonar was placed at the midpoints of the shorter edges, while in Pool C, it was positioned 10 m from the short edge along the longer side. We recruited over 10 volunteers for training data collection in Pool B and 18 subjects for evaluation, with a maximum of 10 concurrent subjects whose heights vary from 173 m to 190 m, weights from 65 kg to 82 kg. In addition, two volunteers who had previous drowning experiences provided valuable suggestions. The volunteers' swimming skills cover novices, skilled amateur swimmers, and professional swimmers, including lifeguards. Note that only skilled swimmers simulated movement, struggling, and drowning. The volunteers performed activities (as detailed in 4.4) to serve as ground truth labels. They were positioned at various distances to simulate different levels of crowdedness, with some volunteers out of line of sight to mimic real-world deployment scenarios. RGB cameras are set up poolside to record their ground truth locations. We conducted 20 training sessions over 18 months (from September 2022 to August 2024), taking around 3 hours each and featuring different activity combinations and localizations, with a professional lifeguard present for safety. The endeavor resulted in the acquisition of 94 hours of training data, encompassing approximately 39,000 sonar images, predominantly featuring sessions with fewer than 5 individuals. The evaluation was carried out in pools A, B and C, yielding a total collection of 310, 722, and 300 images, respectively.

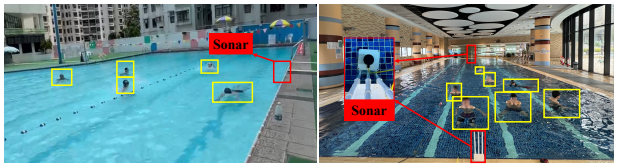


Figure 10: Examples of data collection in two pools.

5.2 Evaluation Metrics and Baselines

We use various quantitative metrics to evaluate the performance of object detection and tracking. We deployed several state-of-the-art methods for activity recognition.
Detection of Human Subjects. We assess object detection using three popular metrics, i.e., F1-score, miss detection rate (MDR), and intersection over union (IoU). F1-score is the harmonic mean of precision and recall rates, which considers both false detection and the miss rate in object detection. MDR describes the ratio of missed bounding boxes to the total number of bounding boxes. IoU represents the ratio of the area of overlay (AoO) to the area of union (AoU) between predicted and ground truth bounding boxes.
Tracking. We use two metrics for evaluating tracking performance, i.e., frame rate per second (FPS) and tracking rate (TR). FPS reflects the sonar scanning speed. TR reflects the performance of tracking by $TR = N_{\text{correct}}/N_{\text{all}}$, where N_{correct} and N_{all} represent the successfully tracked subjects and the total number of subjects, respectively.
Recognition. We use the accuracies of individual activities of five classes and combined activities as evaluation metrics. We combine the moving, motionless, and splashing as the "safe" classes and the rest of the classes (i.e., struggling and drowning) as the "dangerous" classes.
Baselines. We deploy five typical image-denoising baselines to evaluate AquaScan's denoising method: AverageBlur [12], BilateralFilter [54], GaussianBlur [40], MedianBlur[25], and BM3D [9]. Besides, our method is compared with KBNet [65], a state-of-the-art end-to-end image denoising algorithm.

Two baseline methods of image object detection are deployed since it is similar to our problem setting: One is the YOLOv5 [29], while the other is YOLOv10 [57] which has a better precision and latency. We select YOLOv10m and YOLOv5m, which balance latency and accuracy. Except for one-stage object detection, we implement another baseline based on a convolutional recurrent neural network (CRNN) in [10], which can extract spatial and temporal information

Table 2: The values of hyperparameters.

Parameter	Value	Parameter	Value	Parameter	Value
$\bar{F}_{\text{motion}}$	0.95	$T_{\text{struggling}}$	20s	$T_{\text{motionless}}$	60s
R_{min}	1.0	D_{min}	0.3m	D_{max}	0.6m
IoU_{min}	0.5	d_{min}	0.3m	d_{max}	0.6m
(x/y)	1/3				

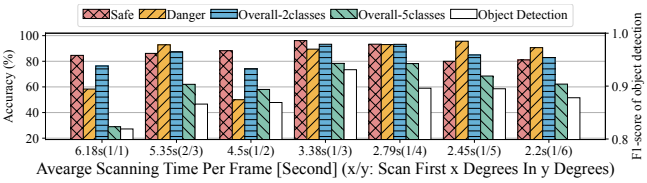


Figure 11: Object detection and activity recognition vs. intermittent scanning parameters.

about human activities. We use 3 consecutive frames as the input for all baselines and AquaScan's recognition models.

5.3 Hyperparameter Settings and Evaluation

Table 2 summarizes AquaScan's hyperparameter setting. $\bar{F}_{motion}$ is set to 0.95 with a time window of 30s. $T_{struggling}$ is set to 20 s and $T_{motionless}$ is set to 60 s according to the guidelines in [8]. $T_{motionless}$ is set as 60 s to show the ability of recognizing quite drowning. The remaining parameters (i.e., R_{min}, and IoU_{max}) were determined based on the collected training data. We evaluate various settings (i.e., x/y) of AquaScan's intermittent scanning in Section 4.2. Fig. 11 illustrates the recognition accuracy of intermittent scanning from full (i.e., 1/1) scan to 1/6 scan under the hyperparameters of table 2. Increasing the number of skipped angles can enhance the frame rate, which improves performance in dangerous scenarios involving significant movement. Specifically, the results indicate that 1/1, 2/3, and 1/2 perform poorly because they cannot maintain the right traces.

Increasing the number of skipped angles shows a decline in recognizing safe activities due to missed object detections, as indicated in the results of 1/3, 1/4. In conclusion, the 1/3 configuration achieves the highest detection accuracies across all the metrics, so AquaScan adopts the 1/3 setting for intermittent scanning.

5.4 An End-to-End Evaluation

To evaluate the end-to-end performance of AquaScan against baselines, we invited ten and five volunteers in Pools A and B at the same time, respectively. Each volunteer conducted one or multiple activities at various locations during each data collection session. Fig. 12 depicts the confusion matrices of classification performance of AquaScan and baselines among five pool activities. AquaScan achieves an accuracy of 91.5% for five-class recognition, detecting 97% of safe and 97% of dangerous activities, outperforming all baselines.

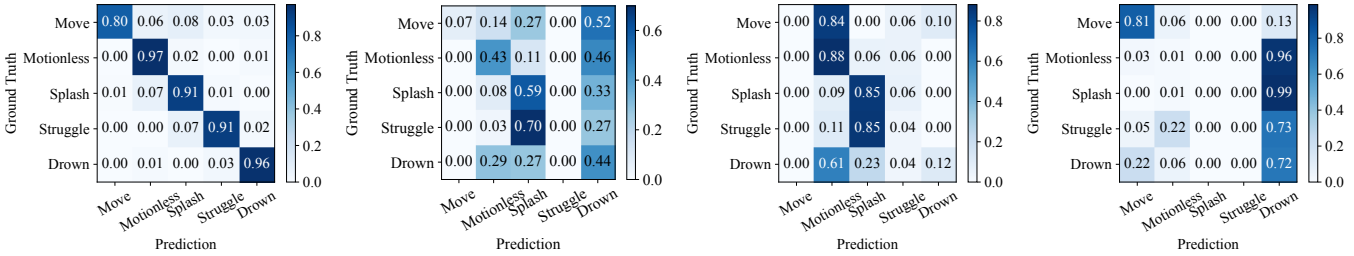
In contrast, baseline methods perform suboptimally: YOLOv5 and YOLOv10 exhibit significantly lower overall performance (35.2%, 49.1%) for five-class recognition due to their inability to extract temporal information, leading to poor performance in swimming, struggling, and drowning. Notably, YOLOv10 recognizes only 13% of dangerous activities. The CRNN achieves an overall accuracy of 31.4% for five-class recognition and 24% and 72% for safe and dangerous activities, respectively. While CRNN excels at extracting temporal features in motionless-related activities and drowning, it cannot extract the spatial and time-domain features as AquaScan does, causing numerous false alarms and reduced practical utility. In summary, YOLOv5, YOLOv10, and CRNN prove less effective for this application due to their limitations in extracting relevant features from sonar images for accurate activity recognition.

We evaluate AquaScan's latency of activity recognition. The average overall latency is 9.18 seconds (s), comprising 2.92s for scanning one sonar image which is used to smooth recognized activities and 1.86s for processing one image. Fig. 13 illustrates the detection latency for five activities[3], revealing that 90% of activities are detected within 14 seconds, 95% within 16 seconds, and all within 30 seconds. According to [8], drowning typically lasts 20 to 60 seconds. Given these timeframes, the observed detection latencies fall within acceptable limits, providing lifeguards with adequate time to be alerted and respond during critical drowning moments. A slightly longer delay occurs when the subject's state switches between motionless and splashing. We observe seldom motionless states that take a longer time to detect the motionless due to the water waves. We also observe very few drowning cases take longer time to detect due to the late detection of other activities. Note that the average detection delays of motionless and drowning cases are 9.86 and 9.13 seconds, respectively. Both the longest and average detection delay are within the detection thresholds set by the guideline [8].

5.5 Evaluation of Image Reconstruction

To evaluate the effectiveness of AquaScan's image reconstruction in Section 4.2, we train AquaScan's recognition model using raw sonar images with incomplete angles. Figs.14(i) and (ii) show that without image reconstruction, AquaScan's detection and recognition accuracy suffer. Fig. 14(i) effectively recognizes dangerous activities due to recovered key features, while Fig. 14(ii) misses subjects, leading to poor tracking and recognition performance. In addition, Figs. 14(iii) and (iv) also show that AquaScan without physical-aware denoising and data augmentation of training data exhibits significant performance degradation in activity recognition, respectively. Unconstructed sonar images lose

[3]Latency calculation: If the subject is detected at t_1, and the activtiy is recognized at time t_2, the latency of recognizing this activity is t_2-t_1. if the activtiy is not recognized and ends at t_3, then the latency is t_3-t_1.



(a) AquaScan (Safe:0.97 Danger:0.97) **(b) YOLOv5 (Safe:0.54 Danger:0.40)** **(c) YOLOv10 (Safe:0.92 Danger:0.13)** **(d) CRNN (Safe:0.24 Danger:0.72)**

Figure 12: Confusion matrics of pool activity recognition using AquaScan and baselines. The safe and dangerous values in the sub-captions show the classification accuracy of safe and dangerous activities, respectively.

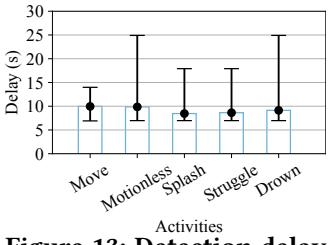


Figure 13: Detection delay across activities.

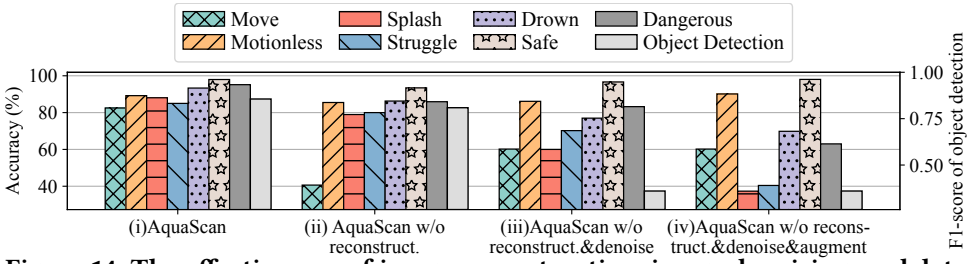


Figure 14: The effectiveness of image reconstruction, image denoising, and data augmentation.

the features of motion, leading to low detection performance of motion-related activities such as splashing and struggling.

5.6 Evaluation of Object Detection

We further test the performance of AquaScan's object detection module against various baselines. We arrange 5 volunteers in the scanning range of one sonar. First, we deploy four classic algorithms and two advanced methods (KBNet [65] and BM3D [9]) to evaluate denoising methods. AverageBlur, BilateralFilter, GaussianBlur, and MedianBlur are four classic algorithms that smooth images by a kernel-wise filter for high-frequency noise removal and edge extraction. BM3D is a popular block-wise algorithm for image noise removal. KBNet is a state-of-the-art learning-based image denoising algorithm that provides a pre-trained model for deployment in our evaluation. Table 3 shows the performance of the baselines and AquaScan. AquaScan surpasses all baselines, yielding the highest F1-score (0.857) and IoU (0.519), with a low miss rate (0.061). While image blur methods reduce noise, they struggle with extremely noisy images. BM3D, optimized for Gaussian noise, underperforms with water-induced reflections. KBNet, designed for RGB images, fails to generalize to sonar data.

We also conducted an ablation study to present the significance of AquaScan's prior physical knowledge. Our system obtains a low F1-score (0.467) without physical prior knowledge since it cannot adaptively denoise the echoes that cause large false detections. We validate the object detection with different crowdedness, where the gap between still humans is $0.5m$,$0.6m$,$0.7m$, and more at $5-6m$ in front of the sonar.

Table 3: Detection performance of baselines, AquaScan, AquaScan without physical-aware noise removal.

Methods	F1-Score	MDR	IoU
AverageBlur [12]	0.404	0.467	0.301
BilateralFilter [54]	0.346	0.551	0.279
GaussianBlur [40]	0.346	0.535	0.273
MedianBlur [25]	0.431	0.437	0.293
KBNet [65]	0.434	0.330	0.313
BM3D [9]	0.509	0.249	0.357
Ours	**0.857**	0.061	**0.519**
Ours w/o PhI	0.467	**0.021**	0.505

AquaScan misses 35.0% of the subjects due to signal overlaps with $0.5m$ gap, detects 96.7% with $0.6m$ and $0.7m$, and 100% subjects with $\geq 0.8m$.

Tracking. We compare the baseline performance (full scan images) and our proposed scanning strategy from the tracking perspective in real-world experiments with 5 subjects. The TR of AquaScan is 95.5% TR, while the naive scan is 75.4%. The FPS of AquaScan is 0.296, while the naive scan is only 0.162. Our method outperforms in both two metrics since a higher sampling rate and F1-score benefit matching detected objects and existing trajectories.

5.7 Impact Factors

Performance Across Various Regions. Distance between swimmers and sonars affects the performance since echoes attenuate through propagation, which decreases the accuracy of activity recognition. Fig. 15 shows the accuracy at different distances (1-5m, 5-8m, 8-11m, and ≥11m) between subjects and sonar. Subjects in far regions suffer from weaker

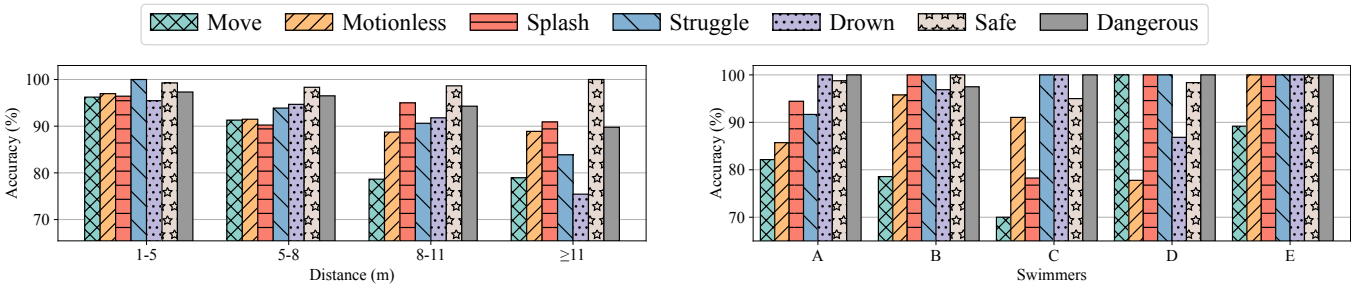


Figure 15: Recognition performance across distances.

Figure 16: Recognition performance across swimmers.

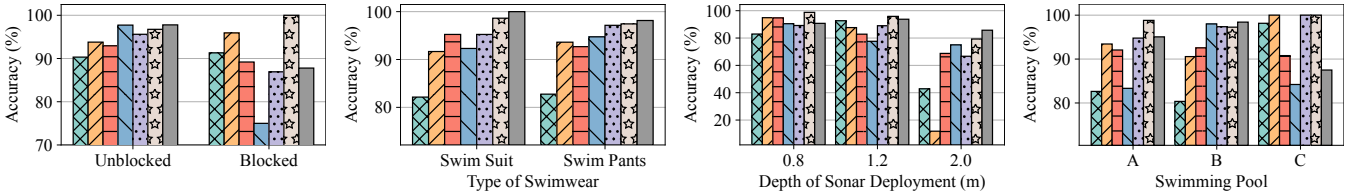


Figure 17: Impact of object blockage.

Figure 18: Impact of swimwear.

Figure 19: Impact of sonar depths.

Figure 20: Impact of different pools.

signals so the accuracy of 5 classes decreases. Dangerous activities observed during a time slot are less affected. Moving, motionless, and splashing are mixed, which does not degrade safe activity recognition.

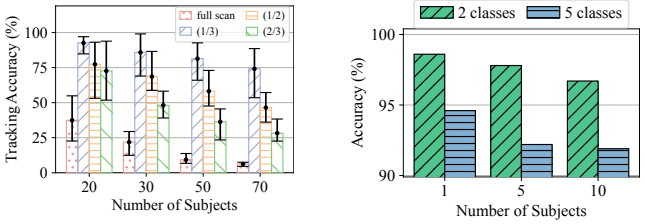


Figure 21: Numeric results of tracking performance.

Figure 22: Classification accuracy across various subject numbers.

Number of Swimmers. We evaluate the impact of different swimmer numbers in the pool through both numerical and real-world experiments. We developed a simulation algorithm that can randomly initialize subject positions and simulate their movement with random speed and direction changes. We simulated the sonars to obtain the locations of subjects asynchronously. In the numerical simulation, we assume each subject matches one existing trajectory according to the location and velocity to judge whether subjects can be paired. Fig. 21 shows the tracking rate (TR) under different x/y and number of subjects. As x/y decreases, TR increases due to more sampling points on subjects' movement trajectories, enhancing accuracy. Even with 133 subjects in a crowded pool, about 54% can be tracked correctly. In addition, we evaluate 2-class and 5-class activity recognition with 1, 5, and 10 subjects in the pool from real-world experiments. Results in Fig. 22 show that accuracy for five classes decreases due to inter-subject interference. The number of subjects has minimal impact on overall 2-class activity monitoring accuracy.

Performance Across Diverse Body Shapes. We invite 5 swimmers with varied body shapes representing the body shapes in our evaluation dataset: A (177cm 78kg), B (177 67kg), C (170cm 65kg), D (182cm 80kg), E (188cm 83kg). The result in Fig. 16 shows that although different users may have diverse performances, AquaScan can still achieve more than 90% accuracy for detecting safe and dangerous activities.

Performance with Occlusion. Occlusion is a major concern in an acoustic sensing system, so we test the performance when the subjects are blocked and interfered (by another volunteer), or unblocked. The accuracy of splashing, struggling and drowning decreases to 89.2%, 75.0%, and 86.9% due to weak reflections from blocked subjects. AquaScan achieves 100% and 87.8% for safe and dangerous activity detection, which indicates it can work for pool monitoring.

We test AquaScan with swimmers wearing full-body and partial swimsuits. One subject performed activities at varying distances in Pool B, both in swimsuits and swim pants. Four additional subjects acted as typical swimmers. Fig. 18 shows that swimsuits minimally impact activity recognition.

Performance with Sonars at Different Depths. We conduct experiments with sonars deployed at 0.8m, 1.2m, and 2.0m depths. Figure 19 shows that the accuracy for safe activities drops to 79.3% due to weak echoes from partially covered subjects. Splashing and struggling cause intense water disturbances, leading to better detection than other activities. What's more, weak echoes decrease the performance of tracking, which causes worse activity performance

with sonars deployed at 2.0m. Hence, it is recommended to deploy the sonar at 0.8m or 1.2m.

Performance in Different Pools. To show the generalization of the swimming pool, we test AquaScan in Pool A and Pool B with concurrent 5 subjects in the scanning range of one sonar. Due to weather and rules, we finally only invited 1 volunteer in Pool C. The results in Fig. 20 indicate that the pool has little influence on the performance of dangerous activities (at least 87.5% in Pool C), which is acceptable for swimming pool monitoring.

Swimming Strokes. To verify that AquaScan can recognize moving subjects with different strokes, we invite 4 volunteers as interferences and 1 swimmer to conduct four breaststrokes, backstrokes, crawl strokes, and butterfly strokes. Performances for 4 strokes are 81.5%, 83.3%, 85.7%, and 84.6%, respectively. None of the moving is misclassified into dangerous activities. Swimmers of breaststroke have relatively low performance since they do not conduct intense motion like the other three strokes. Besides, noise caused by intense motion can degrade the performance of tracking and recognizing moving subjects.

6 DISCUSSION

Multi-Sonar Collaboration for Larger Pools. A multi-sonar strategy for larger areas like 50-meter pools can be adopted when a single sonar's coverage is limited. In future work, we plan to develop a collaborative approach where multiple sonars enhance coverage and detection accuracy through co-denoising and co-localization. By comparing the paths detected by multiple sonars, we can differentiate objects from noise. High cosine similarity and short Euclidean distance between paths can confirm the detection of the same object, improving accuracy in large pool areas.

Safety of ultrasonic sonars. According to the Canadian authority [48], an underwater ultrasonic imaging device should satisfy thermal index (TI)$<$ 1.5; mechanical index (MI)$<$ 1.9, and spatial-peak temporal-average intensity (I_{SPTA})$<$ 720 mW/cm^2 for safety. The scanning sonar adopted by AquaScan uses the power of 5 watts [2]. The thermal indexes of a baby and an adult are 0.73 and 0.35, respectively, which are more than 2 times less than the standard. This means that 1.3 hours and 2.87 hours of continuous exposure are safe for a baby or an adult, respectively. Note that the actual scanning time is only 11% of the duration in each sonar image because of intermittent directional scanning. The mechanical index is 9.14×10^{-5}, which is 20,000 less than the requirement. The scanning area at 1m in front of the sonar is 232 cm^2, so we can calculate I_{SPTA} as 21 mW/cm^2. In summary, the ultrasonic sonar used by AquaScan fully complies with the existing safety regulations.

Maximum Number of Subjects. We assume that each person occupies at least $1.14m^2$ for separate detection (see Section 5.6). In a pool measuring 50 m $\times$ 25 m, we can detect over 1000 swimmers, which is theoretically extremely crowded. Considering blockage, AquaScan can still detect at least 100 subjects in a pool.

Detection for Children in the Pool. Detecting young children poses challenges for AquaScan due to their smaller body sizes. Suppose that a 2-year-old child swims 15 meters from the sonar, their typical waist is 42.4 cm [14], making the body trunk at least 13.5 cm wide and occupying 1.13 grades. With the sonar's horizontal width at 2.22 grads, our (1/3) scanning scheme skips only 0.78 grads, ensuring the baby's trunk is detected. Additionally, the movement of their body and arms generates a larger dataset over time, enhancing detection.

System Deployment Overhead. The AquaScan systems in three pools are actually deployed after the pools are built. The scanning sonars are installed on the pool's edge walls using a stand, which has minimal interference with the existing pool installation.

Customizing Sonars. Customizing a sonar that can access raw data is not cost-effective since processing raw ultrasonic waves in ultra-high frequency requires a dedicated high-speed data processing unit such as FPGA. Commercial sonars' images contain the location and motion features of subjects, which is adequate for pool monitoring. AquaScan is the first-of-its-kind system deploying commercial sonars for underwater human monitoring. In the future, AquaScan can be further developed to recognize more fine-grained and long-distance activities in open water, such as oceans

7 CONCLUSION

In this paper, we propose AquaScan, a novel scanning sonar-based underwater sensing system for human activity monitoring in swimming pools. AquaScan features an innovative intermittent scanning strategy, physical-aware image denoising, and a multi-dimensional feature extraction and state-transition framework for activity recognition. Field tests in public pools demonstrate AquaScan's high accuracy (91.5%) and real-time performance (9.18 seconds), offering an effective solution for aquatic safety and pool management.

8 ACKNOWLEDGEMENTS

This paper is supported in part by the Innovation and Technology Fund of Hong Kong under ITS/231/22FP and the Research Grants Council (RGC) of Hong Kong under GRF 14207123. We are grateful to our shepherd Prof. Fadel Adib and anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] 2023. BlueRobotics. https://bluerobotics.com
[2] 2023. BlueRobotics. https://bluerobotics.com/store/sensors-sonars-cameras/sonar/ping360-sonar-r1-rp/.

[3] 2024. Code for AquaScan. https://github.com/CUHK-AIoT-Sensing/AquaScan.

[4] Gaddi Blumrosen, Ben Fishman, and Yossi Yovel. 2014. Noncontact wideband sonar for human activity detection and classification. *IEEE Sensors Journal* 14, 11 (2014), 4043–4054.

[5] Antoni Burguera and Gabriel Oliver. 2016. High-resolution underwater mapping using side-scan sonar. *PloS one* 11, 1 (2016), e0146396.

[6] Tuochao Chen, Justin Chan, and Shyamnath Gollakota. 2023. Underwater 3D positioning on smart devices. In *ACM SIGCOMM 2023*.

[7] Enrique Coiras, Yvan Petillot, and David M Lane. 2007. Multiresolution 3-D reconstruction from side-scan sonar images. *IEEE Transactions on Image Processing* 16, 2 (2007), 382–390.

[8] American Red Cross. 1995. *Lifeguarding Today*. Mosby Lifeline. https://books.google.com.hk/books?id=VJ7IqLw5A14C

[9] Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. 2011. BM3D frames and variational image deblurring. *IEEE Transactions on image processing* 21, 4 (2011), 1715–1728.

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96. 226–231.

[12] Linwei Fan, Fan Zhang, Hui Fan, and Caiming Zhang. 2019. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art* 2, 1 (2019), 7.

[13] Yihan Feng, Yaoguang Wei, Shuo Sun, Jincun Liu, Dong An, and Jia Wang. 2023. Fish abundance estimation from multi-beam sonar by improved MCNN. *Aquatic Ecology* 57, 4 (2023), 895–911.

[14] Cheryl D Fryar, Margaret D Carroll, Qiuping Gu, Joseph Afful, and Cynthia L Ogden. 2021. Anthropometric reference data for children and adults: United States, 2015-2018. (2021).

[15] .guinnessworldrecords. 2023. The Longest waistline. https://www.guinnessworldrecords.com/world-records/67531-largest-waist

[16] Upulie Handalage, Nisansali Nikapotha, Chanaka Subasinghe, Tereen Prasanga, Thusithanjana Thilakarthna, and Dharshana Kasthurirathna. 2021. Computer vision enabled drowning detection system. In *2021 3rd International Conference on Advancements in Computing (ICAC)*. IEEE, 240–245.

[17] Tim Hansen and Andreas Birk. 2023. Synthetic Scan Formation for Underwater Mapping with Low-Cost Mechanical Scanning Sonars (MSS). *IEEE Access* (2023).

[18] Alex E Hay and Douglas J Wilson. 1994. Rotary sidescan images of nearshore bedform evolution during a storm. *Marine Geology* 119, 1-2 (1994), 57–65.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Lixing He, Haozheng Hou, Zhenyu Yan, and Guoliang Xing. 2022. Demo Abstract: An Underwater Sonar-Based Drowning Detection System. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. 493–494. https://doi.org/10.1109/IPSN54338.2022.00047

[22] Nobuhiro Hiranoa, Kosuke Onishia, and Seiichi Serikawaa. 2016. Suggestion of High Precision Drowning Detection System Using Plural Radio Modules. (2016).

[23] Hiroumi HORIMOTO, MAKI Toshihiro, Kazuya KOFUJI, and Takashi ISHIHARA. 2018. Autonomous sea turtle detection using multi-beam imaging sonar: Toward autonomous tracking. In *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*. IEEE, 1–4.

[24] Haochen Hu, Zhi Sun, and Lu Su. 2020. Underwater motion and activity recognition using acoustic wireless networks. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 1–7.

[25] Thomas Huang, GJTGY Yang, and Greory Tang. 1979. A fast two-dimensional median filtering algorithm. *IEEE transactions on acoustics, speech, and signal processing* 27, 1 (1979), 13–18.

[26] Kohji Iida, Rika Takahashi, Yong Tang, Tohru Mukai, and Masanori Sato. 2006. Observation of marine animals using underwater acoustic camera. *Japanese Journal of Applied Physics* 45, 5S (2006), 4875.

[27] Ulf Jensen, Franziska Prade, and Bjoern M Eskofier. 2013. Classification of kinematic swimming data with emphasis on resource consumption. In *2013 IEEE International Conference on Body Sensor Networks*. IEEE, 1–5.

[28] Jia-Xian Jian and Chuin-Mu Wang. 2021. Deep learning used to recognition swimmers drowning. In *2021 IEEE/ACIS 22nd International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, 111–114.

[29] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. 2022. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo* (2022).

[30] H Paul Johnson and Maryann Helferty. 1990. The geological interpretation of side-scan sonar. *Reviews of Geophysics* 28, 4 (1990), 357–380.

[31] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. 2015. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE international conference on computer vision*. 4696–4704.

[32] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[33] Masahiro Kobayashi, Yuto Omae, Kazuki Sakai, Akira Shionoya, Hirotaka Takahashi, Takuma Akiduki, Kazufumi Nakai, Nobuo Ezaki, Yoshihisa Sakurai, and Chikara Miyaji. 2018. Swimming motion classification for coaching system by using a sensor device. *ICIC Express Letters Part B: Applications* 9, 3 (2018), 209–217.

[34] Hovannes Kulhandjian, Narayanan Ramachandran, Michel Kulhandjian, and Claude D'Amours. 2019. Human Activity Classification in Underwater using Sonar and Deep Learning. In *Proceedings of the International Conference on Underwater Networks & Systems*. 1–5.

[35] Aboli Kulkarni, Kshitij Lakhani, and Shubham Lokhande. 2016. A sensor based low cost drowning detection system for human life safety. In *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 301–306.

[36] Min Li, Houwei Ji, Xiangcun Wang, Liyuan Weng, and Zhenbang Gong. 2013. Underwater object detection and tracking based on multi-beam sonar image processing. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 1071–1076. https://doi.org/10.1109/ROBIO.2013.6739606

[37] Jie Lian, Xu Yuan, Ming Li, and Nian-Feng Tzeng. 2021. Fall Detection via Inaudible Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.

[38] Wen-Hung Liao, Zhung-Xun Liao, and Ming-Je Liu. 2003. Swimming style classification from video sequences. In *16th IPPR Conference on Computer Vision, Graphics and Image Processing, Kinmen, R. OC.*

[39] Qiongzheng Lin, Zhenlin An, and Lei Yang. 2019. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[40] Tony Lindeberg. 2024. Discrete approximations of Gaussian smoothing and Gaussian derivatives. *Journal of Mathematical Imaging and Vision* (2024), 1–42.

[41] J Macqueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press.*

[42] Yann Marcon, Eberhard Kopiske, Tom Leymann, Ulli Spiesecke, Vincent Vittori, Till von Wahl, Paul Wintersteller, Christoph Waldmann, and Gerhard Bohrmann. 2019. A Rotary Sonar for Long-Term Acoustic Monitoring of Deep-Sea Gas Emissions. In *OCEANS 2019 - Marseille*. 1–8. https://doi.org/10.1109/OCEANSE.2019.8867218

[43] Danilo Navarro, Gines Benet, and Milagros Martínez. 2007. Line based robot localization using a rotary sonar. In *2007 IEEE Conference on Emerging Technologies and Factory Automation (EFTA 2007)*. IEEE, 896–899.

[44] Yuto Omae, Yoshihisa Kon, Masahiro Kobayashi, Kazuki Sakai, Akira Shionoya, Hirotaka Takahashi, Takuma Akiduki, Kazufumi Nakai, Nobuo Ezaki, Yoshihisa Sakurai, et al. 2017. Swimming style classification based on ensemble learning and adaptive feature value by using inertial measurement unit. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 21, 4 (2017), 616–631.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[46] A Pozo-Ruz, JL Martínez, and A García-Cerezo. 1997. Integration of a rotary sonar in the mobile robot RAM-2. *IFAC Proceedings Volumes* 30, 7 (1997), 143–147.

[47] Muhammad Ramdhan, Muhammad Ali, Samura Ali, MY Kamaludin, et al. 2018. An early drowning detection system for internet of things (iot) applications. *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 16, 4 (2018), 1870–1876.

[48] Defence Research and Development Canada. [n. d.]. The safety of diver exposure to ultrasonic imaging sonars. https://publications.gc.ca/collections/collection_2018/rddc-drdc/D68-11-10-2018-eng.pdf.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[50] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 474–485.

[51] David C Schwebel, Heather N Jones, Erika Holder, and Francesca Marciani. 2010. Lifeguards: A forgotten aspect of drowning prevention. *Journal of injury and violence research* 2, 1 (2010), 1.

[52] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 591–605.

[53] Deividas Tarasevičius and Artūras Serackis. 2020. Deep learning model for sensor based swimming style recognition. In *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. IEEE, 1–4.

[54] Carlo Tomasi and Roberto Manduchi. 1998. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 839–846.

[55] Xiaofeng Tong, Lingyu Duan, Changsheng Xu, Qi Tian, and Hanqing Lu. 2006. Local motion analysis and its application in video based swimming style recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 2. IEEE, 1258–1261.

[56] Mario Vittone. 2010. Drowning doesn't look like drowning. *Repéré à http://mariovittone. com/2010/05/154* (2010).

[57] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458* (2024).

[58] Wikipedia. 2023. List of tallest people — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/List_of_tallest_people

[59] Haesang Yang, Sung-Hoon Byun, Keunhwa Lee, Youngmin Choo, and Kookhyun Kim. 2020. Underwater acoustic research trends with machine learning: Active SONAR applications. *Journal of Ocean Engineering and Technology* 34, 4 (2020), 277–284.

[60] Jing Yang, James P Wilson, and Shalabh Gupta. 2019. Diver gesture recognition using deep learning for underwater human-robot interaction. In *Oceans 2019 Mts/Ieee Seattle*. IEEE, 1–5.

[61] Qiang Yang and Yuanqing Zheng. [n. d.]. Neural Enhanced Underwater SOS Detection. *Power (dB)* 80, 60 ([n. d.]), 40.

[62] Qiang Yang and Yuanqing Zheng. 2023. AquaHelper: Underwater SOS Transmission and Detection in Swimming Pools. (2023).

[63] Zhijian Yang, Xiaoran Fan, Volkan Isler, and Hyun Soo Park. 2022. PoseKernelLifter: Metric Lifting of 3D Human Pose using Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13179–13189.

[64] Chi Zhang, Xiaoguang Li, and Fei Lei. 2015. A novel camera-based drowning detection algorithm. In *Chinese Conference on Image and Graphics Technologies*. Springer, 224–233.

[65] Yi Zhang, Dasong Li, Xiaoyu Shi, Dailan He, Kangning Song, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. 2023. Kbnet: Kernel basis network for image restoration. *arXiv preprint arXiv:2303.02881* (2023).