Sensors and Materials, Vol. 38, No. 1 (2026) 295–310
MYU Tokyo

S & M 4300

AI-based Object Recognition and Risk Detection Technology for Swimming Pool Safety Prediction

Sung-Sam Hong,[1,2] Hyungjin Jeon,[2] Chanlim Park,[2] and Hwayoung Kim[3*]

[1]Department of Information Security, Hankyung National University, Pyeongtaek-si 17738, Korea
[2]Department of AI Lab, Smart Safety Laboratory Co., Ltd., Seongnam-si 13494, Korea
[3]Department of Maritime Transportation, Mokpo National Maritime University, Mokpo-si 58628, Korea

(Received December 4, 2025; accepted January 16, 2026)

Keywords: drowning detection, AI, safety, risk detection, object recognition

Drowning incidents in swimming pools remain a critical public health issue globally, where rapid detection and response significantly impact survival rates. Traditional human-based surveillance and sensor-based systems face challenges of cognitive limitations, environmental constraints, and high costs. We propose a two-stage (Two-Stage) framework that, utilizing video streams captured by camera sensors, employs computer vision and deep learning to precisely detect human objects in real time and subsequently classify 'risk' behaviors to predict safety incidents. The core focus of this study is (1) to benchmark the performance of the latest real-time object detection models, YOLOv12 and RT-DETR, for the 'person' detection module (Stage 1) in a pool environment and (2) to validate the efficacy of a hybrid data strategy—integrating public datasets (Public-Sets) with a custom-collected dataset (Custom-Set)—to optimize this detector's performance against the known challenge of data scarcity. Experiments were conducted in three scenarios (public data only, custom data only, and a hybrid combination). The results revealed a stark trade-off between detection speed and accuracy; RT-DETR-R50 demonstrated exceptional real-time speeds (approximately 140 FPS), whereas YOLOv12-L provided superior accuracy (*mAP*) but was not viable for real-time use. We also found that the Public-Set (from Roboflow (9500 images) produced the highest general accuracy (*mAP@*.5), while the Custom-Set (1,986 images) produced the highest localization precision (*mAP@*.5:.95). Through this research, an empirical foundation for the 'detection' component (Stage 1) of the proposed framework was established and the path for integration with Stage 2 'precise behavior classification' models in future work was outlined.

1. Introduction

Drowning is a leading cause of preventable death worldwide, accounting for 7% of all injury-related fatalities and ranking as the third leading cause of unintentional injury death.(1) It is a severe public health problem, particularly for children aged 1–14.(2) The World Health

*Corresponding author: e-mail: hwayoung@mmu.ac.kr
https://doi.org/10.18494/SAM6101

ISSN 0914-4935
https://myukk.org/

Organization (WHO) reported that more than 236,000 people die from drowning annually,[1] representing not only a tragic loss of life but also a significant economic burden.

The primary line of defense against these incidents, particularly in supervised pools, has traditionally been visual surveillance by trained lifeguards.[3] However, this method has inherent limitations. A recent report from Royal Life Saving Australia indicated that when lifeguard-to-patron ratios exceed 1:75, the detection of a drowning incident is significantly delayed.[4] This suggests that human visual attention is highly vulnerable to cognitive fatigue from sustained monitoring and the challenges of visually cluttered scenes.[4]

To mitigate this human limitation, various technological approaches have been proposed, which can be broadly categorized as sensor-based and computer vision-based.[2] Sensor-based systems include wearable devices (e.g., specialized goggles) or the installation of underwater sonar[5] or other sensors.[6] This approach has the advantage of being relatively free from privacy concerns.[7] However, it introduces problems such as forcing users to wear equipment, the need for periodic charging and maintenance,[1] and the limited underwater communication range.[7] Fixed sensor systems like sonar, in particular, have a very high initial installation cost and system complexity,[8] making them difficult to deploy in public or rural pools.

In contrast, computer vision (CV)-based systems utilize surveillance cameras (CCTV) installed in and around the pool, with AI utilized for analyzing the footage in real time to detect dangerous situations.[9] This method allows for nonintrusive monitoring without requiring user-worn devices, offering great potential in terms of scalability. However, CV-based approaches also face serious technical challenges that need to be resolved to enable practical application.[6]

For a CV-based drowning detection system to operate effectively in a real-world environment, it must overcome the following complex challenges:[6]

- Environmental Factors: The swimming pool is an extremely harsh environment for CV models. Irregular light reflections on the water's surface, splashing from swimmers, water turbidity,[10] and rapid changes in indoor or outdoor lighting conditions are major factors that severely degrade the accuracy of object detection algorithms.
- Data Scarcity: The performance of deep learning models is dictated by the quality and quantity of training data. However, large-scale, high-quality public datasets depicting actual drowning incidents are virtually nonexistent owing to ethical issues and the difficulty of collection.[11] This has been the biggest bottleneck in training a model to learn real-world risk situations.
- High False Alarm Rate: Compounding the data scarcity problem, the initial response to drowning often looks visually similar to 'normal' playful behavior, such as vigorous splashing.[6] This causes the system to frequently misidentify normal activity as a risk (False Positive),[6] which degrades the system's reliability. Data imbalance[9] also biases the model toward the majority (normal) class.
- Privacy Concerns: The use of cameras for continuous video monitoring, especially in a space like a swimming pool where swimwear is worn, could raise serious privacy issues.[9] This has been a significant barrier to the social acceptance of the technology.

In this study, we focus primarily on the challenges of data scarcity and real-time detection accuracy in CV-based swimming pool surveillance systems. Although a two-stage AI

framework—comprising swimmer detection (Stage 1) and risk behavior classification (Stage 2)—is conceptually proposed, the experimental scope of this paper is explicitly limited to the design, implementation, and validation of the Stage 1 "person" detection module processing video data captured by camera sensors. The performance of Stage 2 behavior classification models will be discussed in future work rather than being experimentally evaluated in this study. This paper is focused on the development and validation of the *Stage 1 'person' detection* model. The specific academic and technical contributions of this research are as follows.

- Real-time Detection Model Benchmarking: We applied the latest attention-centric real-time detection model (YOLOv12) and a Transformer-based real-time detection model (RT-DETR)[9] to the "swimming pool safety" domain for the Stage 1 person detection module. We quantitatively compared the real-time performance (detection time) and detection accuracy (*mAP*) of both models to identify the optimal architecture for this specific domain.[9]
- Proposal of a 2-stage Risk Classification Framework: We proposed a 2-stage framework that rapidly detects 'person' objects in Stage 1 and classifies their behavior as 'Normal' or 'Risk' in Stage 2. We experimentally validated the performance of Stage 1.
- Hybrid Dataset Strategy Validation: To maximize the performance of the Stage 1 detector, we empirically verified the impact of a 'hybrid training strategy'—in which public datasets are combined with a custom-simulated dataset—on model generalization and 'person' detection accuracy (*mAP*) through a systematic ablation study.

2. Related Work

2.1 Object detection in aquatic and underwater environments

Object detection in aquatic and underwater environments presents fundamentally different challenges from standard terrestrial object detection. In a recent comprehensive survey on AI-based underwater object detection (UOD),[10] optical distortion from light absorption and scattering, blurriness from water turbidity,[12] and irregular, rapidly changing illumination were named as the core challenges of UOD.

Early drowning detection research has relied on traditional machine learning (ML) algorithms[10] or classic image processing techniques like background subtraction and contour detection.[1] However, these methods are highly vulnerable to the aforementioned environmental factors and lack robustness.

In recent years, the rapid advancement of deep learning (DL) technology, particularly convolutional neural networks (CNNs), has changed the paradigm. In many modern drowning detection systems, the use of DL models such as YOLO, Faster R-CNN, and ResNet[12] to detect human objects and analyze their behavior has been attempted. However, the performance of these DL-based systems is entirely dependent on the quality and quantity of training data,[13] and most prior research faced a common limitation: an absolute shortage of real-world drowning data.[8]

2.2 Real-time object detection models: CNN/attention vs Transformer

As requested, in this study, we selected two state-of-the-art (SOTA) model architectures suitable for the time-critical task of swimming pool surveillance, YOLOv12 and RT-DETR, as candidates for the Stage 1 detector.

- YOLOv12 (Attention-centric)[14]
  The You Only Look Once (YOLO) series has long been synonymous with 1-stage object detectors,[15] consistently offering an excellent balance of speed and accuracy. YOLOv12, announced in February 2025, diverges from previous CNN-focused frameworks by integrating 'attention-centric' mechanisms to maximize performance. This model was designed to leverage the performance benefits of attention mechanisms while maintaining the speed of CNN-based models. YOLOv12 is already being evaluated for applications in real-time video surveillance and human activity recognition. Figure 1 shows the architectures of YOLOv12.
- RT-DETR (Transformer-based)
  DETR (detection Transformer) was the first model to successfully introduce the Transformer architecture to object detection, creating a complete end-to-end pipeline that did not require complex post-processing (like NMS).[15] Real-time DETR (RT-DETR)[16] has maintained this structural advantage while being equipped with an efficient hybrid encoder and IoU-aware query selection[15] to achieve real-time performance. Transformer-based models show strength in understanding global image context via self-attention mechanisms.[17] This characteristic holds potential for more robustly distinguishing objects from complex, cluttered backgrounds (e.g., water splashes and other swimmers) in a pool environment.[17] Its applicability has already been explored in aquatic[17] and coastal debris[18] environments. Figure 2 shows the overview of TR-DETR.
- Performance Debate (YOLO vs DETR)
  The comparison of the performances of these two architectures is a current topic of academic debate. In the 2024 CVPR paper "DETRs Beat YOLOs on Real-time Object Detection",[15] it was reported that on the standard COCO benchmark, the RT-DETR-R50 model achieved both higher accuracy (AP) and faster speed (FPS) than the YOLOv8-L model.[15] However, this claim did not apply universally to all domains. For example, a 2023 Korean study on

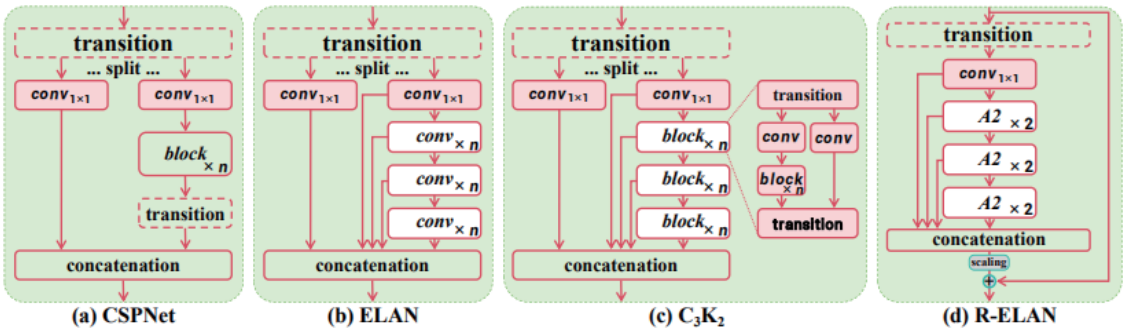

Fig. 1. (Color online) Architectures of popular modules (YOLOv12).

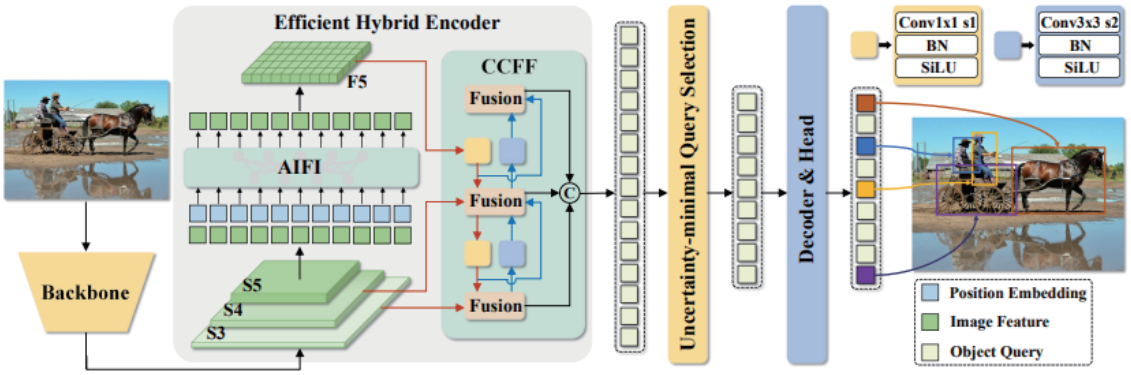

Fig. 2. (Color online) Overview of RT-DETR.

coastal debris detection[18] reached the opposite conclusion: YOLOv8 consistently outperformed RT-DETR in both *mAP* and FPS.[18] Similar findings[17] suggest that model superiority is domain-dependent.

YOLOv12 extends conventional one-stage detectors by incorporating optimized attention mechanisms into the backbone and neck structures, allowing enhanced feature aggregation under complex visual conditions such as water reflections and partial occlusions. RT-DETR, in contrast, adopts a Transformer-based end-to-end formulation that removes the need for post-processing steps such as non-maximum suppression (NMS), enabling more stable inference latency. These architectural differences motivate their comparative evaluation in time-critical swimming pool safety applications.

2.3 Drowning behavior classification and dataset limitations

The goal of the 2-stage framework proposed in this study was to move beyond simple person detection to accurately distinguish between 'normal' swimming and 'abnormal' drowning behaviors. Influential research specifically defined the typical behavioral patterns of a drowning person.[19] These included (1) the instinctive drowning response (IDR), (2) a "climbing ladder motion", and (3) a "backward stroke".[19]

To accurately detect these 'risk' behaviors (positive samples), a model must inevitably be trained on a wide variety of 'safe' behaviors as 'negative samples'.[7] These negative samples must include not only standard strokes but also idling[13] and even vigorous splashing.

The problem was the absence of a standardized, large-scale benchmark dataset that encompassed these diverse and fine-grained behaviors. Currently accessible public datasets might focus only on specific environments (e.g., underwater views[20]) or include limited behavioral scenarios. Consequently, many prior studies, like the one proposed here, required the construction of custom, simulated datasets, which supported the necessity of our 'custom-collected dataset' methodology.

3. Two-stage Framework for Risk Prediction to Enhance Swimming Pool Safety

We proposed a two-stage approach to achieve both real-time performance and, ultimately, classification accuracy in the swimming pool environment. Stage 1 was the 'detection stage,' where a high-speed object detector identified all individuals in the pool. Stage 2 was the 'classification stage,' where the detected objects were classified as 'normal' or 'risk'. Figure 3 shows the structure of proposed frameworks. This paper is focused on the design and experimental validation of the Stage 1 'swimmer detection' model.

3.1 Proposed framework overview

The proposed end-to-end system architecture consists of the following four modules:

1. Input: Real-time video streams (N frames per second) are received from multi-viewpoint CCTV cameras in the pool area.
2. Stage 1: Swimmer Detection: A pretrained, real-time object detection model (verified in this study: YOLOv12 or RT-DETR) processes each incoming frame (t). The output of this stage is the bounding box coordinates $\lceil x, y, z, h \rceil$ for all detected 'swimmer' objects within the image.
3. Stage 2: Risk Classification (Proposed)
   - The bounding box area for each 'person' detected in Stage 1 is passed as input to a Stage 2 classification model.
   - This Stage 2 model (e.g., a spatiotemporal model[21]) analyzes the temporal movement or pose[11] of the object to classify its behavior as 'normal' or 'risk'. (This stage was not part of this paper's experimental scope.)
4. Output and Alerting: If a specific object is classified as 'risk' by the Stage 2 model for k consecutive frames (e.g., 3 s), it is considered a real risk situation, not a temporary false alarm. The system then immediately transmits an alert to lifeguards or managers via smartwatches or alarm systems,[1] prompting swift intervention.

3.2 Stage 1: Swimmer Detection Module Design

The Stage 1 ('person' detection) component is the core element determining the speed and efficiency of the entire system. If this detector misses a person (False Negative) or detects the

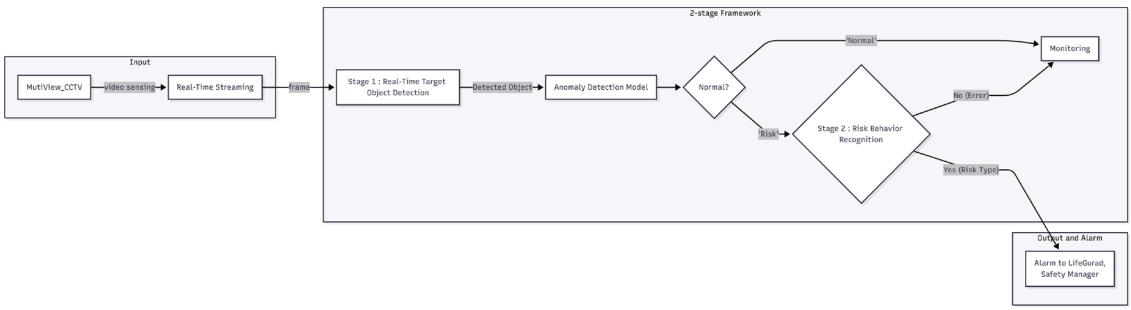

Fig. 3. Structure of proposed framework.

wrong area (False Positive), the Stage 2 classification model cannot even begin to function. Therefore, we focus on validating the Stage 1 detector's performance, setting two SOTA models as candidates.

- Candidate Model 1: YOLOv12
  - YOLOv12 is based on the 'attention-centric' architecture. This design incorporates optimized attention modules to improve object recognition without unnecessarily increasing computational cost (latency).
  - A key challenge in the pool environment is the varying scale of objects and partial occlusion from splashes or other swimmers.[12]
  - YOLOv12's attention mechanism is designed to efficiently process key features within this complex visual information, allowing it to robustly detect 'person' objects of various scales.
- Candidate Model 2: RT-DETR
  - RT-DETR is an end-to-end Transformer-based model offering the advantage of a simplified pipeline by eliminating the need for post-processing steps like NMS.[15]
  - The self-attention mechanism[22] of its hybrid encoder considers the global context of the entire image.[23] This provides the potential to more clearly distinguish the target object ('person') from complex, noisy backgrounds (e.g., waves and light reflections).[17]

4. Datasets and Experimental Design

We designed a systematic experiment to analyze how the performance of the Stage 1 'Swimmer Detection' model is affected by the composition of the training dataset.

4.1 Public dataset aggregation and reprocessing

To ensure the model's generalization performance, we collected, refined, and consolidated several publicly accessible datasets to build a Public-Set.

- Key Included Datasets:

1. Drowning Detection Dataset: A self-made dataset published on GitHub by Wang[24] was utilized in a 2024 study on improved YOLOv5 algorithms. It contains 8,572 images simulating drowning and swimming positions from drone perspectives.
2. Roboflow Datasets: A collection of datasets aggregated from the Roboflow Universe by searching for the 'swim' keyword.[25] This includes several open-source projects focused on detecting 'swimming', 'drowning', or 'person' objects in various pool environments.

- Data Reprocessing (Single-Class Consolidation): A total of 9,500 images were aggregated from these public sources for the ‹Public-Set›. The goal of this Stage 1 detection experiment was *not* to classify risk, but to detect the 'person' object itself. Therefore, for this experiment, all related class labels (e.g., ‹swimming', 'struggling', 'drowning', and 'idle') were remapped and consolidated into a single 'person' class.

4.2 Custom dataset curation methodology

To address the domain mismatch problem and develop a detector specialized for the actual target environment, a ‘Custom-Set’ was constructed.

- Data Acquisition:

1. Location: Data was collected at an indoor swimming pool.
2. Participants: One professional swimmer simulated both normal swimming and abnormal behaviors.
3. Filming: A 1080p camera was used to capture video from surface and underwater perspectives.

- Data Processing and Annotation:

1. Unique images were extracted from the video at a rate of 1 frame per second.
2. This process yielded a total of 1986 unique images for the ‹Custom-Set›.
3. To fit the purpose of this Stage 1 detection experiment, professional annotators performed bounding box labeling (YOLO/COCO format) for all human objects in every image using the single ‘person’ class.
4. No data augmentation techniques were applied to this dataset.

All video data used for the custom dataset were collected in compliance with local ethical guidelines and privacy regulations. Informed consent was obtained from all participants prior to data collection, and the recorded data were used solely for research purposes. No personally identifiable information was retained, and all annotations were limited to bounding box coordinates without facial or biometric identification.

4.3 Experimental scenarios

We designed an ablation study with the following three scenarios to systematically analyze the effect of dataset composition on the Stage 1 ‘person’ detector’s performance. This design was based on the results of prior research indicating that mixing real and simulated data was beneficial to model performance. Each of the three datasets was split into training, validation, and test partitions in a 60/20/20 ratio. The specific image counts for each split are detailed in Table 1.

- Scenario 1: Public-Set Only
  - Training*:* Models (YOLOv12 and RT-DETR) were trained *only* on the ‘Public-Set’ (9,500 images) (single ‘person’ class).
  - Hypothesis: The model should have basic detection capabilities but might suffer from domain mismatch.

Table 1
Dataset splits for training, validation, and testing.

Dataset	Train	Val	Test	Total
Hybrid set	6892	2297	2297	11486
Custom dataset	1192	397	397	1986
Public dataset	5700	1900	1900	9500

- Scenario 2: Custom-Set Only
  - Training: Models were trained *only* on the 'Custom-Set' (1,986 images) (single 'person' class).
  - Hypothesis: The model should show high performance on the test set because it is highly specialized for the target domain,[13] but there are the risk of overfitting and poor generalization.
- Scenario 3: Hybrid-Set (Public + Custom)
  - Training: The 'Public-Set' (9500 images) and 'Custom-Set' (1986 images) were combined. Models were trained on the resulting 'Hybrid-Set' (11486 images) (single 'person' class).
  - Hypothesis: By learning from both the broad diversity of the Public-Set and the domain specificity of the Custom-Set, this model should achieve the best overall performance.
- Evaluation Metrics:
  - Object Detection Performance: *mAP*@0.5, *mAP*@0.5:0.95.[18]
  - Real-time Performance: Detection time (s).

5. Experimental Results and Analysis

5.1 Stage 1: Comparison of 'person' object detection performance

All experiments were conducted on an NVIDIA RTX 3090 Ti. For each of the three scenarios, the YOLOv12-m and RT-DETR models were trained, validated, and tested using the respective 60/20/20 dataset splits defined in Table 1. For example, the models in Scenario 1 were trained on 5700 images and tested on 1900 images, while models in Scenario 3 were trained on 6892 images and tested on 2297 images. The performance of each model against its corresponding test set is detailed in Table 2.

5.2 Analysis of experimental results

First, the results of the comparative analysis of the models (YOLOv12 vs RT-DETR) are given below. Figures 4 and 5 show the detection visualization results of each model using the public dataset.

- Speed (Detection Time): The experimental results indicate that both RT-DETR-R50 and YOLOv12 achieved inference times within the real-time operational range. RT-DETR-R50

Table 2
Performance of 'person' object detection for each training scenario and model.

Training Scenario	Model	*mAP*@.5 (Accuracy)	*mAP*@.5:.95 (Precision)	Detection time (ms)
Scenario 1 (Public-Set)	YOLOv12	0.9777	0.7432	5.01
	RT-DETR	0.9730	0.6930	7.2
Scenario 2 (Custom-Set)	YOLOv12	0.9717	0.8704	4.82
	RT-DETR	0.9290	0.7890	7.1
Scenario 3 (Hybrid-Set)	YOLOv12	0.9725	0.7493	5.09
	RT-DETR	0.9680	0.7090	6.9

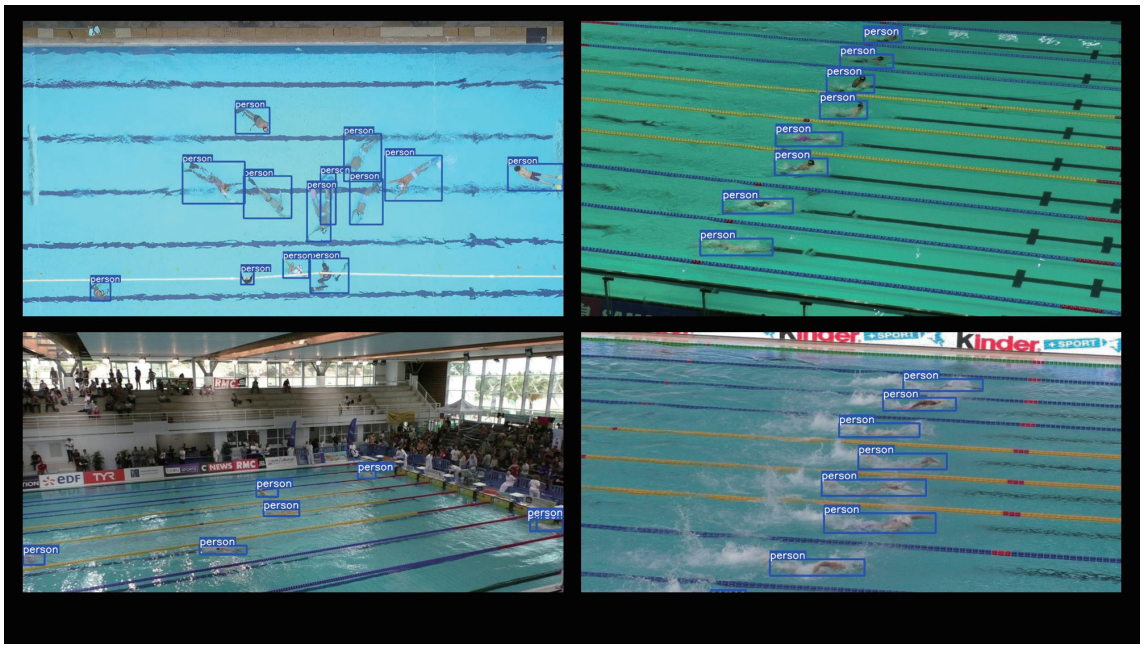

Fig. 4. (Color online) YOLOv12 detection result (public dataset).

Fig. 5. (Color online) RT-DETR detection result (public dataset).

processed frames at approximately 7.2 ms (≈139 FPS), while YOLOv12 achieved a shorter inference time of about 5.1 ms (≈196 FPS). These findings confirm that both models are suitable for real-time detection and analysis in swimming pool surveillance applications.

- Accuracy (*mAP@*.5): In terms of general accuracy (a loose IoU threshold), YOLOv12 consistently outperformed RT-DETR in all three scenarios. The highest *mAP@*.5 (0.9777) was achieved by YOLOv12 trained on the Public-Set (Scenario 1).
- Precision (*mAP@*.5:.95): For strict localization precision, YOLOv12 was again superior in all scenarios. It achieved its highest performance (0.8704) when trained *only* on the Custom-Set (Scenario 2). This suggested that the custom dataset, though small, contained very high-quality and precise bounding box annotations.

Second, the analysis of the results for each of the dataset configuration scenarios is as follows.

- The hypothesis that the Hybrid-Set (Scenario 3) would yield the best performance was not supported by the results. Instead, the dataset strategy presented a clear trade-off.
- Scenario 1 (Public-Set Only): This scenario produced the highest overall *general accuracy* (*mAP@*.5 of 0.9777) with the YOLOv12 model. This indicated that the large, diverse public dataset was excellent for training the model to detect the *presence* of a person.
- Scenario 2 (Custom-Set Only): This scenario produced the highest *precision* (*mAP@*.5:.95 of 0.8704) with the YOLOv12 model. This significant lead in precision, despite the small dataset size (1986 images), strongly implied that the custom data's annotations were of exceptionally high quality, resulting in training the model for superior *localization*. Figures 6 and 7 show the detection visualization results of each model using the public dataset.
- Scenario 3 (Hybrid-Set): This scenario seemed to present a compromise but did not top any category. Its performance (e.g., 0.9725 *mAP@*.5) was slightly lower than the Public-Set in general accuracy and significantly lower than the Custom-Set in precision. Figures 8 and 9 show the detection visualization results of each model using the public dataset.

Lastly, a significant trade-off exists between speed and accuracy. RT-DETR-R50 is the only model suitable for real-time (140+ FPS) applications. YOLOv12 is vastly superior in both general accuracy (*mAP@*.5) and localization precision (*mAP@*.5:.95), but its ~5 s detection time makes it completely unsuitable for real-time use. It would only be viable for offline, post-event analysis. Dataset strategy is also key: the Public-Set was best for general detection, while the Custom-Set was superior for high-precision localization.

6. Conclusions

6.1 Research summary and conclusion

We proposed a 2-stage AI framework for ensuring swimming pool safety and empirically validated the performance of the Stage 1 'person' detection module. To accomplish this, we benchmarked the latest deep learning detectors, YOLOv12 and RT-DETR, on a single-class 'person' detection task.

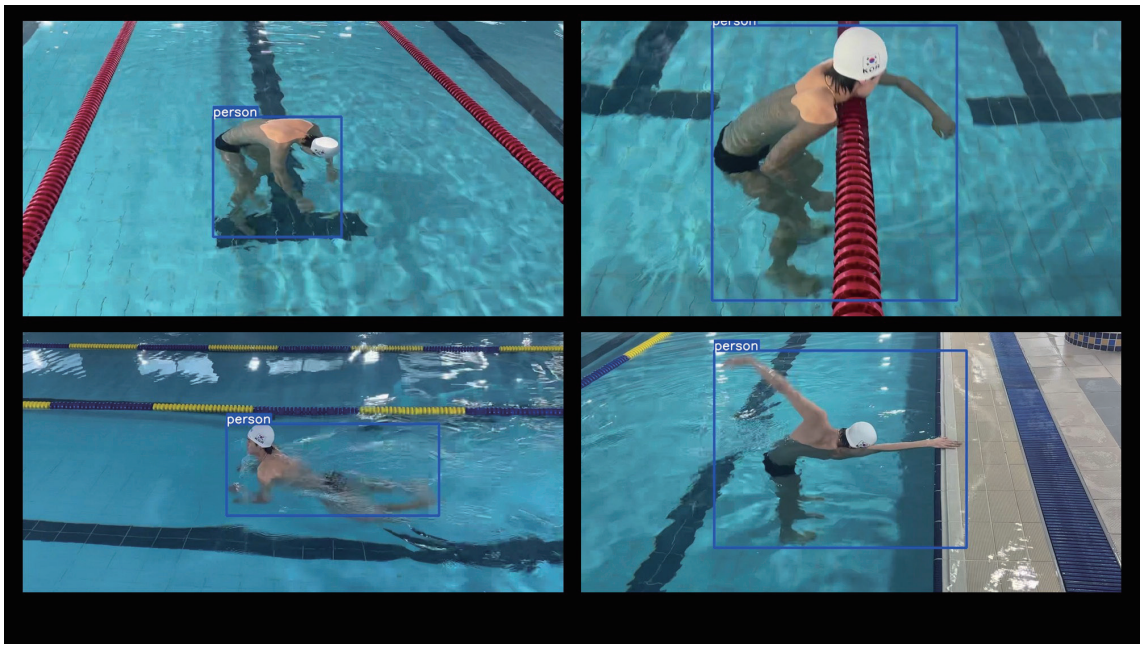

Fig. 6. (Color online) YOLOv12 detection result (custom dataset).

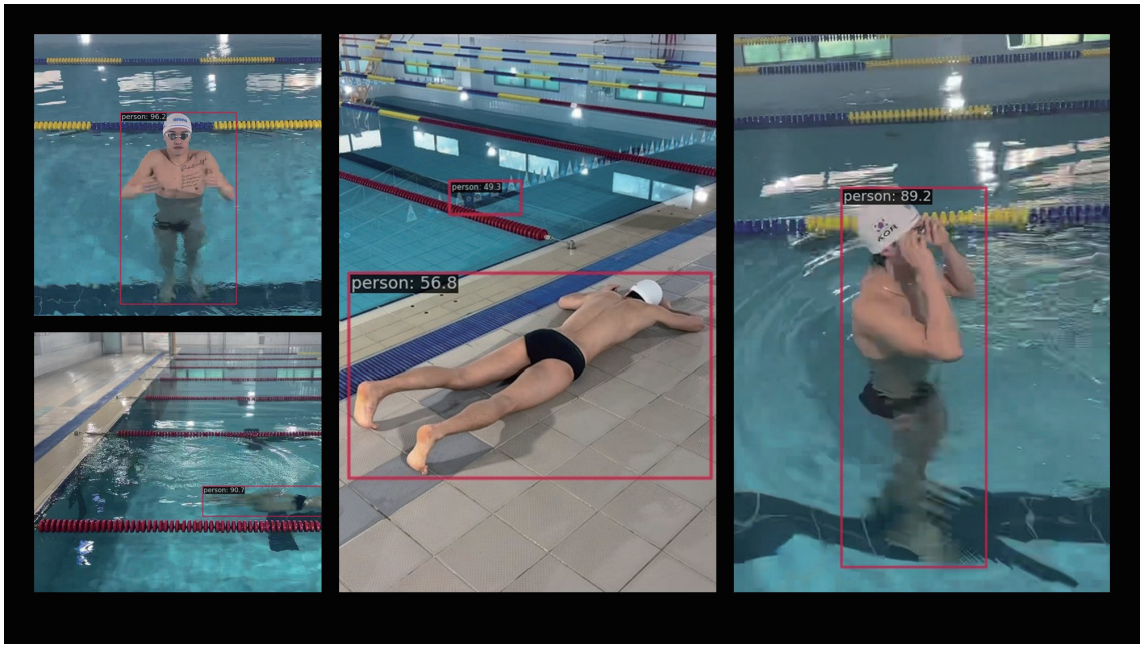

Fig. 7. (Color online) RT-DETR detection result (custom dataset).

The main conclusions were as follows.

- Sensor-Based Monitoring: This study successfully established a robust model for object detection and risk analysis by utilizing the video data collected through camera sensors in a swimming pool environment. The integration of high-resolution visual sensors with deep learning architectures proved to be an effective foundation for automated safety surveillance.

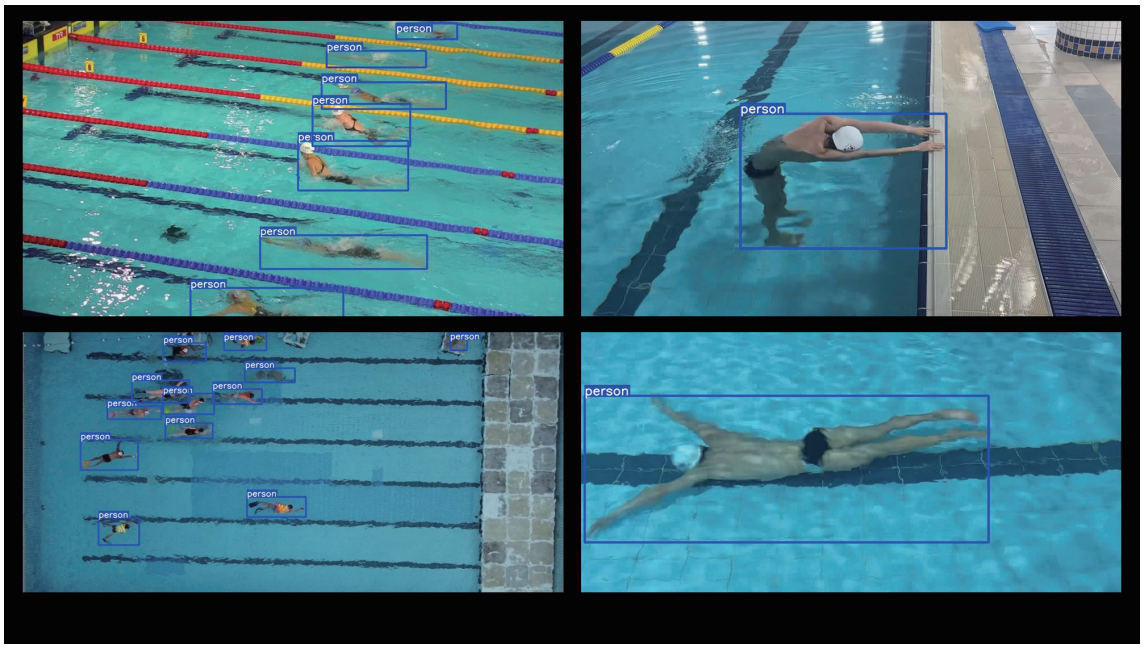

Fig. 8. (Color online) YOLOv12 detection result (hybrid dataset).

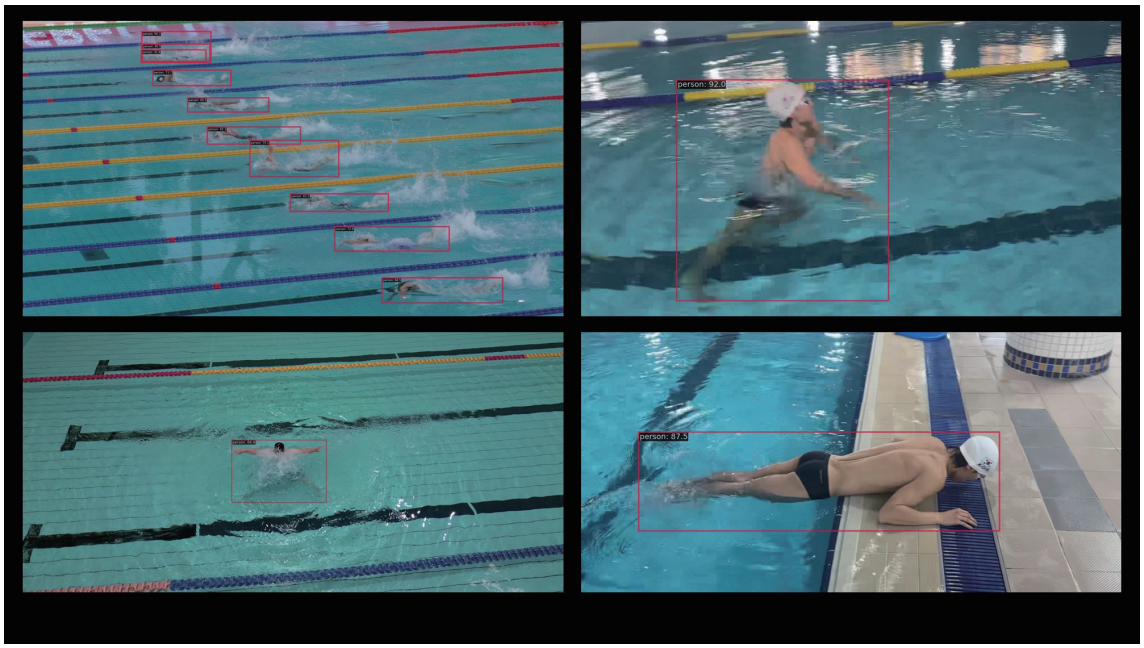

Fig. 9. (Color online) RT-DETR detection result (hybrid dataset).

- Model Performance: A critical trade-off between speed and accuracy was identified. RT-DETR-R50 was the only real-time viable model (approximately 140 FPS). YOLOv12-L, while extremely slow (approximately 0.2 FPS), provided far superior accuracy in both general (*mAP@*.5) and precise (*mAP@*.5:.95) metrics.

- Data Strategy: The hypothesis that the Hybrid-Set would be optimal was not supported. Instead, the Public-Set (9,500 images) proved to be best for general accuracy (0.9777 *mAP@*.5 with YOLOv12-L), while the Custom-Set (1,986 images) was best for high-precision localization (0.8704 *mAP@*.5:.95 with YOLOv12-L).
- 1-Step Performance: For a *real-time* framework, the RT-DETR-R50 trained on the Public-Set (Scenario 1) provides the best balance of high speed (0.0072s) and high general accuracy (0.9730 *mAP@*.5). For *offline analysis* where accuracy is paramount, YOLOv12-L trained on the Custom-Set (Scenario 2) provides the best precision.
- Real-Time System: From a real-time performance perspective, both RT-DETR-R50 and YOLOv12 demonstrated inference speeds sufficient for deployment in time-critical swimming pool safety systems. With detection times of 7.2 and 5.1 ms per frame, respectively, the experimental results confirm that the proposed Stage 1 detection module can operate in real time, enabling continuous monitoring and timely risk assessment within the envisioned two-stage safety framework.

6.2 Limitations of the study

We successfully validated the Stage 1 'person' detection model, but it had clear limitations.

- Detection-Only Study: This research was focused *only* on Stage 1 'person' detection. While the model succeeded in finding the *location* of swimmers rapidly and accurately, the core task of classifying whether that person's behavior was 'normal' or 'risk' (Stage 2) was not performed.
- Custom Dataset Limitations: The 'Custom-Set' (1,986 images) was effective for improving precision but was collected from a single professional swimmer in a single indoor pool. This lacks the diversity to build a robust *behavior classification* model for Stage 2.

6.3 Future works

We successfully validated the Stage 1 'person' detection model. Future research must now focus on implementing the 'Stage 2 precise risk behavior classification' module proposed in Sect. 3.1, on the basis of this validated detector.

- Dataset Re-labeling for Stage 2 Classification:
  Future work must first expand and re-label the custom dataset. Instead of the Stage 1 single-class 'person', it must be annotated for 2-class classification: 'Normal' and 'Risk'.
  - Positive Samples ('Risk'): This class must include behaviors defined in prior research,[19] such as instinctive drowning response (IDR), struggling, and climbing ladder motion.[19]
  - Negative Samples ('Normal'): This class is critical for reducing false alarms.[6] It must therefore include not only common swimming behaviors[7] but also intentionally include abundant 'Hard Negative Samples'—behaviors that look visually similar to 'Risk'—such as vigorous splashing, breath-holding dives, and vertical entry dives.[19]
- Application of Stage 2 Classification Models:
  This stage must take the bounding box of the 'person' from Stage 1 (RT-DETR) as input. To

overcome the limitations of static frames, it must incorporate 'Temporal Context' using spatiotemporal models.[11]

- Proposal 1: Pose Estimation + Time-Series Classification: Apply a high-speed pose estimator (e.g., RTMPose[26]) within the detected bounding box to extract 2D/3D skeleton keypoints. The resulting time-series (sequence) data of these keypoints would be fed into a lightweight time-series classifier (e.g., LSTM, GRU[21], or LightGBM[26]) to classify the temporal patterns of 'normal swimming' vs 'risk'.[27]
- Proposal 2: End-to-End Spatiotemporal Models: Directly combine the Stage 1 detector with a 3D-CNN[28] or LSTM.[21] For instance, a YOLO-LSTM[21] architecture would pass the spatial features extracted by YOLO to an LSTM to model temporal dependences.[21] A 3D-CNN+LSTM combination[28] is a powerful structure already validated in analogous safety fields, such as 'Fall Detection',[28] where analyzing sudden changes in motion is critical.

Acknowledgments

This research was supported by a grant (RS-2025-02472998) provided by Customized Life Safety R&D Project for Public Needs (Phase 2) funded by the Ministry of Interior and Safety (MOIS, Korea).

References

1 J. T. Mathew: Drowning Detection System (2023). https://doi.org/10.13140/RG.2.2.14154.54723
2 M. Shatnawi, F. Albreiki, A. Alkhoori, and M. Alhebshi: Information **14** (2023) 52. https://doi.org/10.3390/info14010052
3 Royal Life Saving Society: https://www.royallifesaving.com.au/about/news-and-updates/news/2024/oct/ai-enhanced-drowning-detection-systems (accessed August 11, 2025).
4 M. Shatnawi, F. A. Albreiki, A. Alkhoori, M. Alhebshi, and A. Shatnawi: Preprints (2024). https://doi.org/10.20944/preprints202410.1058.v1
5 W. C. Kao, Y. L. Fan, F. R. Hsu, C. Y. Shen, and L. D. Liao: Heliyon **10** (2024) e35484. https://doi.org/10.1016/j.heliyon.2024.e35484
6 M. Shatnawi, F. Albreiki, A. Alkhoori, M. Alhebshi, and A. Shatnawi: Information **15** (2024) 721. https://doi.org/10.3390/info15110721
7 S. Jalalifar, A. Belford, E. Erfani, A. Razmjou, R. Abbassi, M. Mohseni-Dargah, and M. Asadnia: Sensors **24** (2024) 331. https://doi.org/10.3390/s24020331
8 M. Akbar and K. Osama: Int. J. Eng. Res. Technol. **13** (2025) 6.
9 T. T. Ünlü, A. L. Mahamat, and M. Turan: WSEAS Trans. Inf. Sci. Appl. **22** (2025) 234. https://doi.org/10.37394/23209.2025.22.20
10 L. Chen, Y. Huang, J. Dong, Q. Xu, S. Kwong, H. Lu, and C. Li: arXiv (2024). https://arxiv.org/abs/2410.05577
11 T. Liu, X. He, L. He, and F. Yuan: IET Image Process. **17** (2023) 1905.
12 S. Bak, H. M. Kim, Y. Kim, I. Lee, M. Park, S. Oh, T. Y. Kim, and S. W. Jang: Korean J. Remote Sens. **39** (2023) 1195.
13 G. Hung and I. F. Rodriguez: arXiv. https://arxiv.org/abs/2509.12682 (2025).
14 M. Shatnawi, F. Albreiki, A. Alkhoori, and M. Alhebshi: Information **14** (2023) 52. https://doi.org/10.3390/info14010052
15 Y. Tian, Q. Ye, and D. Doermann: arXiv. https://arxiv.org/abs/2502.12524 (2025)
16 Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, and J. Chen: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2024) 16965–16974.
17 N. Nemati: arXiv (2025). https://arxiv.org/abs/2510.07346 (2025)
18 B. Sun, H. Tang, L. Gao, K. Bi, and J. Wen: J. Mar. Sci. Eng. **13** (2025) 996. https://doi.org/10.3390/jmse13050996

19 Q. Song, B. Yao, Y. Xue, and S. Ji: Sensors **24** (2024) 6955. https://doi.org/10.3390/s24216955
20 H. Alzaabi, S. Alzaabi, and S. Kohail: Underwater Drowning Detection Dataset (Version 2). https://doi.org/10.6084/m9.figshare.29497235.v2
21 M. Elnady and H. E. Abdelmunim: Sci. Rep. **15** (2025) 17036. https://doi.org/10.1038/s41598-025-01898-z
22 lyuwenyu: https://github.com/lyuwenyu/RT-DETR (accessed April 5, 2025).
23 H. Lv, K. Si, and W. Li: DGP-DETR: A Real-Time Detection Algorithm for Underwater Target Detection (2025).
24 K. Wang: https://github.com/Wang-Kaikai/drowning-detection-dataset (accessed April 5, 2025).
25 Roboflow Universe: https://universe.roboflow.com/search?q=swim (accessed April 20, 2025).
26 T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, and K. Chen: arXiv. https://arxiv.org/abs/2303.07399 (2023)
27 L. Powell, S. Polsley, D. Casey, and T. Hammond: Int. J. Aquat. Res. Educ. **14** (2023) 6.
28 N. Lu, Y. Wu, L. Feng, and J. Song: IEEE J. Biomed. Health Inform. **23** (2018) 314.

About the Authors

Sung-Sam Hong received his B.S. degree from Gachon University, Korea, in 2009 and his M.S. and Ph.D. degrees from Gachon University, Korea, in 2011 and 2016, respectively. From 2016 to 2022, he was a research professor at Gachon University, Korea. Since 2025, he has been an instructor at Hankyung University, Korea. His research interests are in AI, data modeling, cyber security, security and safety system, and AIoT. (sungsamhong@hknu.ac.kr)

Hyungjin Jeon received his B.S. degree from Gangnam University, Korea, in 2022. He has been a vision engineer at SSL Co., Ltd., Korea, since 2024, where he currently serves as an assistant manager. His professional experience includes developing object detection models for road-driving environments and abnormal behavior detection systems for onboard ship monitoring. His research interests include object detection, image segmentation, and pose estimation. (wjswps@smartsafety.co.kr)

Chanlim Park received his B.S. degree from Pennsylvania State University, U.S., in 2013 and his M.S. degree from the Seoul National University, South Korea, in 2013. From 2020 to 2025, he was a researcher at Seoul National University as a Ph.D. candidate. In 2020, he founded Smart Safety Laboratory (SSL) and has since worked as CEO. His research interests are in AI data quality, smart safety, and standardization. (charlie@smartsafety.co.kr)

Hwayoung Kim received his B.S. degree from Mokpo National Maritime University, Korea, in 1998, his M.S. degree from the Mokpo National Maritime University, Korea, in 2002, and his Ph.D. degree from Kyushu University, Japan, in 2007. From 2014 to 2025, he was an associate professor at Mokpo National Maritime University, Korea. Since 2025, he has been a professor at Mokpo National Maritime University. His research interests are in logistics and maritime information systems. (hwayoung@mmu.ac.kr)