





Enhancing Drowning Surveillance with a Hybrid Vision Transformer Model: A Deep Learning Approach

Yingying Zhang¹, Yancheng Li², Qiang Qu³, Huai Lin², Dewen Seng^{2*}

¹ College of Entrepreneurship, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China

² School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

³ Chemical, Mineral and Petroleum Laboratory, Zhanjiang Customs Technology Center, Zhanjiang 524022, China

Corresponding Author Email: sengdw@hdu.edu.cn

Copyright: ©2023 IIETA. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.400647>

ABSTRACT

Received: 8 July 2023

Revised: 26 October 2023

Accepted: 3 November 2023

Available online: 30 December 2023

Keywords:

drowning surveillance, ViT, deep learning, CNN, machine learning

Annually, drowning claims the lives of approximately 372,200 individuals worldwide, averaging 40 fatalities per hour. In response, various technological advancements have been explored, including deep learning-based video and image processing, and wearable devices integrated with human pulse sensors and Light emitting diode (LED)/Liquid-crystal display (LCD) technologies. Despite these efforts, existing solutions have yet to fully address the challenge of accurate drowning detection. This study introduces a novel approach, leveraging a hybrid model that combines a traditional Vision Transformer (ViT) with plain Convolutional Neural Networks (CNNs). This model demonstrates a notable accuracy of 91.5% on a specialized dataset comprising 14,736 images of swimming and drowning scenarios, surpassing conventional methods in efficiency and size. In contrast to larger models like Swin-B, which comprises 88M parameters and achieves a marginally higher accuracy of 92.3%, the proposed model maintains high performance with only 5.9M parameters. The model's development involved pre-training on the ImageNet1K dataset, followed by fine-tuning using the specifically curated local dataset. The resultant system offers a cost-effective, efficient, and compact solution for drowning detection, suitable for various applications. This advancement in drowning surveillance technology highlights the potential of integrating ViT with CNNs in creating effective, resource-efficient models for critical real-world applications.

1. INTRODUCTION

The persistent challenge in drowning detection has been acknowledged as an unsolved issue in public safety [1], predominantly due to the prohibitive costs associated with existing methodologies. The primary objective of this study is the development of a cost-effective and feasible solution. This paper introduces a novel method characterized by its reduced parameter size of 5.9M and an accuracy rate of 91.5% on a locally sourced dataset. This method employs a deep learning approach, specifically utilizing the ViT model [2]. Prior to delving into the specifics of this approach, it is imperative to evaluate the limitations and benefits of the currently prevalent methods in drowning detection.

In this case, two primary categories of methods are identified: wearable equipment-based approaches and computer vision-based techniques. Wearable equipment methods, while boasting high accuracy (near 100%) and a low rate of false positives due to their mechanical nature, are significantly hindered by their cost implications. The efficacy of this approach is contingent upon each swimmer in a pool being equipped with the requisite electronic device, leading to a cost escalation proportional to the number of swimmers. Conversely, computer vision methods, while less financially

burdensome in terms of equipment, incur substantial computational costs, primarily due to the need for substantial Graphics Processing Unit (GPU) resources.

Wearable equipment methods guarantee virtually risk-free swimming environments provided that all individuals in the pool comply with the equipment usage [3]. This method's reliance on individual compliance poses a logistical challenge, alongside the escalating costs with increasing user numbers. The study aims to address these concerns by proposing a method that mitigates the cost and resource limitations of existing approaches.

The advent of computer vision methods, a relatively recent development in the field, predominantly leverages machine learning and deep learning techniques for drowning detection. This includes approaches like Gaussian mixture models [4], as well as deep learning-based methods such as Mask R-CNN [5] and BR-YOLOv4 [6]. A notable shift towards computer vision methods in drowning detection research is observed. An analysis via Google Scholar, using 'drowning detection' as a keyword with a filter for publications post-2023, yielded 1,870 results, with over 1,000 relating specifically to computer vision approaches. This trend underscores the increasing prominence of deep learning in the domain of drowning surveillance. Two primary factors contribute to the

ascendancy of deep learning in drowning detection. Firstly, the inherent efficacy and versatility of computer vision as a tool are undeniable. Secondly, the field of computer vision is marked by rapid advancements, with new and more potent methodologies being reported regularly. Such continual progress offers substantial opportunities for enhancing the effectiveness, speed, and precision of drowning detection systems.

The benefits of employing computer vision methods in drowning detection are twofold. One significant advantage is cost-effectiveness. The expense associated with deploying computer vision-based systems is primarily dependent on the size of the swimming area and the number of pools, rather than the number of swimmers. This feature ensures that costs do not escalate disproportionately with increased user numbers. Furthermore, computer vision methods offer modular flexibility, allowing for algorithmic or model adjustments contingent on the capabilities of the existing hardware infrastructure. In stark contrast, wearable equipment methods lack such adaptability, as any modifications post-manufacture are not feasible.

However, the application of computer vision methods is not without its drawbacks, which can be broadly categorized into two areas. The first pertains to the escalating costs and complexity associated with advanced GPUs. The increasing integration and sophistication of GPUs, coupled with the rising complexity and parameter count of contemporary models, present significant financial challenges. These factors often preclude laboratories, companies, and individuals from training their own models due to budget constraints in acquiring advanced GPUs. The second limitation lies in the accuracy of these methods. Even the most advanced models in the field of computer vision, such as the Swin-L model, struggle to surpass a 95% accuracy threshold, with most models' accuracy ranging between 90% and 94%. In the context of drowning detection, where accuracy equates to human lives, even a minor shortfall in detection capability can have grave consequences. This concern underpins the critical need for improvements in this domain.

This paper presents three key contributions to the field of drowning detection. Firstly, a hybrid ViT model is introduced, which requires a low computational budget and achieves an accuracy of 91.5% on a locally curated dataset. Secondly, the paper details a novel approach named Multiple Windows Surveillance Drowning Detection (MWSDD). This method employs multiple surveillance windows to monitor a single area, enhancing the accuracy of the detection system. This concept parallels the principle of distributed computing, where multiple low-powered units combine to form a highly capable system. Further elaboration on this method is provided in Section 3. Thirdly, a strategic integration of human oversight with the drowning detection system is proposed. This synergy aims to mitigate the risks associated with potential lapses in automated detection, thereby enhancing overall safety and reliability.

2. RELATED WORKS

2.1 ViT

The ViT was adapted from the Transformer architecture, initially proposed by Vaswani et al. for the field of natural language processing [7]. The original ViT maintains the fundamental architecture of the Transformer, with

modifications made primarily to accommodate image inputs. In this adaptation, images are segmented into fixed-size patches, each linearly embedded and augmented with positional embeddings. This process is executed using a single-layer CNN equipped with a fixed-size, learnable convolution kernel. Despite these modifications, the core components of the ViT architecture closely mirror those of the original Transformer. However, the integration of a learnable CNN network in ViT has been associated with potential training instabilities, as elaborated in literature [8].

ViT has been recognized for its effectiveness and utility in various computer vision tasks, with several variants emerging as state-of-the-art models. Notwithstanding their capabilities, these advanced models often entail substantial resource demands, particularly in terms of time, computational power, and GPU memory. For instance, training a ViT-huge model from scratch necessitates approximately 2,500 TPUv3-core-days for pre-training, rendering it prohibitively expensive for most research laboratories. To mitigate the high resource requirements of ViT models, this study references the approach used in LeViT [9], incorporating plain CNNs and pooling layers. This adjustment reduces the number of transformer blocks in the ViT model, thereby decreasing both the parameter count and model size while maintaining stable accuracy. Furthermore, patch projection layers in the ViT are substituted with plain CNNs and pooling layers, enhancing the model's pre-training stability [8].

Figure 1 illustrates a comparative analysis of the structural differences among the original ViT, MoCo v3, and the proposed model.

It is observed that the increasing complexity and size of deep learning models present significant financial barriers, limiting their accessibility to well-funded organizations and research entities. As deep learning technology advances, the trend towards larger models with increasing parameter sizes continues, necessitating the development of more accessible and efficient alternatives.

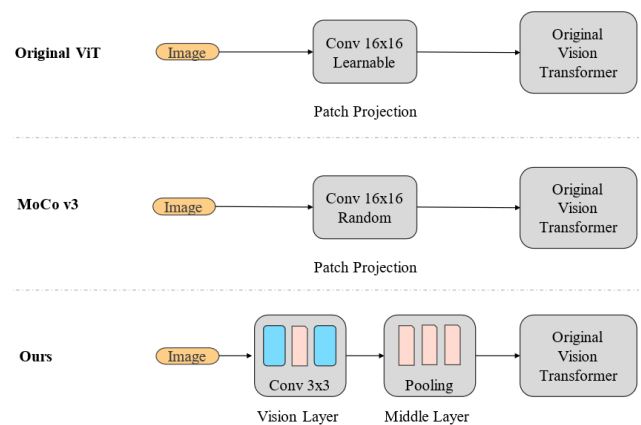


Figure 1. Model structure comparison between original ViT, MoCo v3 and our model

2.2 MoCo v3

MoCo v3, a variant of the ViT model, is documented in literature [8]. This model, based on ViT-L, has achieved an accuracy of 84.1% on ImageNet-1K and has contributed significantly to the understanding of training instabilities in ViT models. The integration of MoCo v3 principles is aimed at enhancing both the stability and accuracy of the proposed model. The fundamental issue in the stability of ViT models

has been identified as the learnable patch projection layer. MoCo v3 addresses this by replacing the learnable layer with a fixed, random patch projection layer. To elaborate, the patch projection layer in question is a CNN layer with a $768 \times 16 \times 16$ convolution kernel. This kernel is capable of compressing a 16×16 image into a singular patch while increasing the channel layers to 768. The hypothesis posited is that the substantial size of the learnable convolution kernel contributes to the training instability. Drawing inspiration from MoCo v3, a modification is introduced to the ViT models by segmenting the large kernel into smaller units. This adaptation has not only resolved the stability issues but has also resulted in an incremental improvement in accuracy.

Figure 1 presents a detailed comparison between the structural elements of the proposed model and other existing models, highlighting the modifications and their implications in terms of stability and performance.

2.3 LeViT

The LeViT model, another variant of the ViT, is explored in literature [9]. It stands out for achieving an accuracy of 87.6% on ImageNet-Real. The structure of the LeViT model is particularly notable; it integrates additional convolutional layers before the ViT segment. This integration has been observed to significantly enhance accuracy, especially in smaller-scale models or scenarios where transformer blocks have limited data for learning. The functionality of the LeViT model aligns closely with the objectives of this research, which prioritizes the development of a compact model that maintains relatively high accuracy.

2.4 Related models

Various methods in the realm of deep learning for computer

vision have been explored in papers [10, 11], each presenting different approaches. A commonality among these methods is their primary focus on improving accuracy. However, the focus of this research diverges from this trend, emphasizing the necessity of finding a balance between cost and accuracy. The aim is to develop a model that not only achieves high accuracy but also remains accessible and affordable, addressing a gap in the current landscape of drowning detection technologies.

3. PROBLEMS AND METHODS

3.1 Problems

In the initial phase of this research, two primary challenges were encountered. The first challenge involved reducing operational costs while maintaining, or ideally improving, the accuracy of the drowning detection system. The second challenge was to establish a reliable contingency plan for situations where the detection system might fail, ensuring timely rescue of individuals at risk of drowning. Solutions to these challenges have been systematically addressed: the first challenge is resolved through the methodologies detailed in Sections 3.2 and 3.3, while the second challenge is addressed in Section 3.4.

3.2 Method and model

In contrast to existing methods, an image classification approach, as opposed to object detection, was employed in this study. This decision broadened the range of applicable models, offering advantages in terms of speed, training efficiency, and potentially higher accuracy.

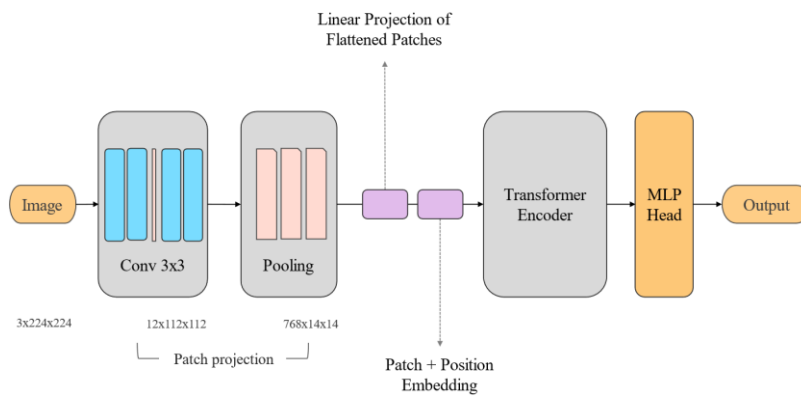


Figure 2. The main structure of our model

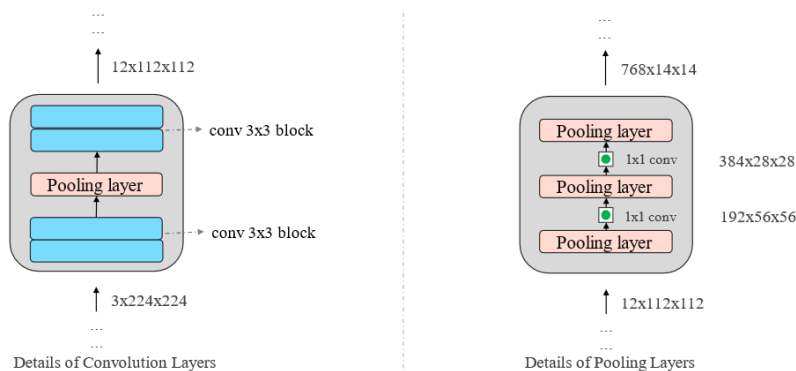


Figure 3. Details of convolution layers and pooling layers

A thorough comparison with prevalent backbone models such as Swin Transformer, DeiT, R-50, and R-100 was conducted. The original ViT model was ultimately selected as the backbone for its simplicity and efficiency. The Swin Transformer, though robust, was deemed excessively large and inflexible for scale adjustments. DeiT's complexity and size exceeded the desired compactness of the study's model. While the traditional CNNs like R-50 demonstrated commendable performance, a comparative analysis revealed that a pre-trained R-50 achieved 91.6% accuracy with 27M parameters. In contrast, the proposed model, inspired by MoCo v3 and LeViT, attained a comparable 91.5% accuracy with only 5.9M parameters, signifying a substantial reduction in computational resource requirements.

The model's architecture is influenced by MoCo v3, wherein four smaller convolution layers and three pooling layers were integrated to replace the original $768 \times 16 \times 16$ convolution kernel. This modification was aimed at resolving ViT's training instability and enhancing accuracy. Additionally, LeViT inspired the detailed design of these convolution layers. The concept involved bundling two convolution layers into a single unit, resulting in two bundled layers. Each of these layers contains two convolution layers devoid of residual links [12], equipped with learnable 3×3 convolution kernels. To maintain the input size stability, padding was utilized during convolution computations. Further specifics of this layered structure can be found in Figures 2 and 3. Integrating this convolution layer structure into the original ViT significantly improved accuracy for smaller ViT models. These models exhibit faster training and response times with only a slight compromise in accuracy, aligning with the project's objectives.

3.3 MWSDD

The implementation of the MWSDD system prompted a critical examination of potential blind spots in surveillance coverage. A notable issue identified was the presence of blind splits at the borders between two adjacent monitoring areas when utilizing a single camera per area. These splits could potentially result in the MWSDD system failing to detect swimmers in distress if they were located precisely at these border areas. To address this challenge and enhance the system's reliability, an overlapping strategy was adopted. Specifically, a 50% overlap of the visual field between neighboring cameras was implemented, effectively eliminating blind splits. This approach not only rectifies the initial flaw but also establishes a double-check security mechanism, significantly improving the accuracy of the MWSDD system. For a detailed illustration of this overlapping technique and its benefits, refer to Figures 4 and 5. It is important to note that increasing the overlap area further could yield higher accuracy, albeit at an increased cost.

Figure 4 presents a comparative visualization of surveillance coverage in a swimming pool. The left image depicts the original swimming pool without any camera surveillance, serving as a baseline for comparison. The right image illustrates the same pool monitored by 16 cameras. Here, the distribution of the cameras and their individual fields of vision are delineated, clearly highlighting the blind splits located at the borders of each camera's monitoring range. These blind splits are critical areas where the MWSDD system's effectiveness is compromised, as they represent zones potentially missed by the surveillance network, MWSDD

system, showcasing a significant advancement in drowning detection surveillance.

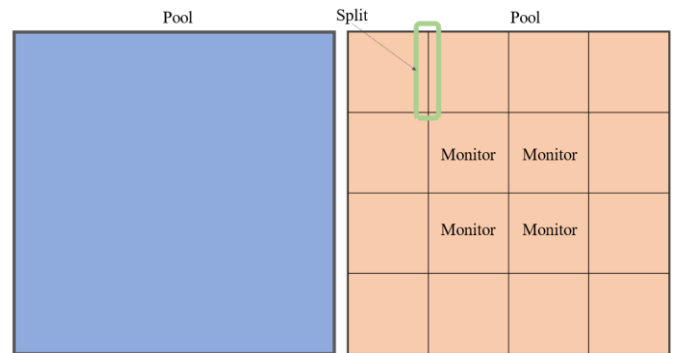


Figure 4. The blind zone of cameras without MWSDD mechanism

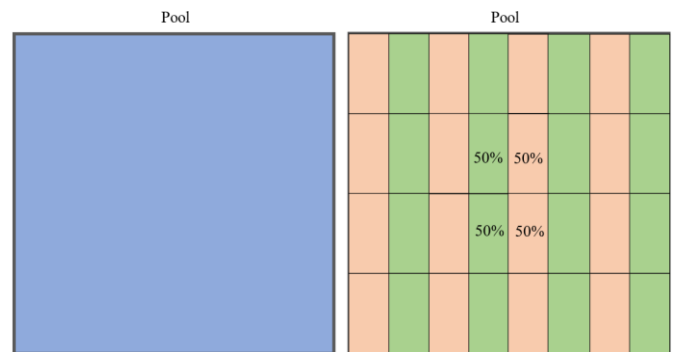


Figure 5. Details of MWSDD mechanism

Figure 5 provides a visual comparison to demonstrate the enhancement in surveillance coverage. Left image depicts the original swimming pool without surveillance for baseline reference. The right image shows the same pool equipped with 32 cameras, each designed to have a 50% overlap in the field of vision with the camera to its left. This configuration effectively addresses the issue of blind splits, as illustrated in Figure 4, by ensuring continuous and comprehensive monitoring coverage across the pool. The overlapping strategy depicted here is integral to the improved functionality and accuracy of the MWSDD system.

3.4 Combination of system and human

This section addresses a critical limitation of the MWSDD system: its inability to achieve 100% accuracy. It is recognized that reliance solely on the MWSDD system could result in missed detections, potentially leading to fatal drowning incidents, an outcome that is imperative to avoid. To mitigate this risk, an integrated approach combining the MWSDD system with human oversight has been developed.

The proposed strategy involves a multi-tiered alarm process, supplemented by human surveillance. Initially, when the MWSDD system identifies a potential drowning incident, an alarm is triggered, and a designated staff member is tasked with immediately alerting a lifeguard. In instances where the system briefly detects a potential drowning but then returns to a normal state, the staff member is required to conduct a thorough review of all monitored areas to confirm the absence of any ongoing distress. Additionally, as a routine precaution, all monitors are to be systematically checked by the staff at three-minute intervals. This integrated approach of

technological surveillance and human vigilance significantly reduces the likelihood of drowning incidents. The implementation of this system requires only two staff members per swimming pool to maintain continuous and effective monitoring.

4. EXPERIMENT

4.1 Experiment setup

The specifics of the hardware setup utilized in this study are delineated in Table 1. The experiment was conducted using the timm PyTorch library [13] as the primary codebase. For accessing pre-training datasets, the PyTorch Datasets library was employed.

Table 1. Hardware implementation

Components	Implementations
CPU	AMD Ryzen 7 5800X 8-Core Processor
GPU	NVIDIA GeForce RTX 3070 8GB
MEMORY	32GB

4.2 Datasets and metrics

Medium-scale image datasets were selected for the pre-training of models. These include ILSVRC-2012 (ImageNet-1k), Oxford-IIIT-Pets [14], CIFAR-100 [15], Oxford Flowers-102 [16]. The decision to utilize medium-scale datasets, as opposed to larger ones, was based on a balance between dataset size and pre-training time. Larger datasets, such as ImageNet-21k and JFM, were deemed prohibitively time-

consuming for the purposes of this research. ImageNet-1k comprises approximately 1.3 million training images across 1000 object categories, while ImageNet-21k contains around 14 million images in approximately 21,000 distinct object categories [17].

Table 2. Swimming and drowning datasets details

Classes	Size	Split Strategies
Total	14,736	80% test, 20% validate
Swimming	7,235	80% test, 20% validate
Drowning	387	80% test, 20% validate
Out of pool	4,351	80% test, 20% validate
Diving	2,763	80% test, 20% validate

For dataset division, the methodology outlined in literature [13] was adhered to. The primary metric for evaluating model performance is top-1 classification accuracy. For the fine-tuning phase, a local dataset consisting of roughly 14,736 images depicting various activities such as swimming, drowning, poolside scenes, and diving was curated. Following standard dataset splitting strategies, 80% of this dataset was used for training and 20% for validation (see Table 2).

4.3 Model selection

The focus of this section is the selection and comparative analysis of different models. Table 3 displays the chosen models along with their top-1 accuracy on pre-training datasets. The model selected for this study is ViT-Ti16, incorporating the modifications detailed in Section 3. All models were pre-trained using images resized to 224×224 pixels.

Table 3. Configurations, param and top-1 accuracy on ImageNet-1k

Model	Param.	ImageNet-1k Top-1 acc. %	Oxford-IIIT-Pets	CIFAR-100	Oxford Flowers-102
ViT-Ti16	5.8M	67.7	88.2	86.2	98.3
ViT-B16	86M	76.6	95.8	91.9	99.6
R-50	27M	77.3	91.1	86.1	94.0
Swin-T [18]	29M	79.4	94.3	95.8	99.2
Swin-B [18]	88M	81.2	97.2	97.0	99.7
DeiT-S [19]	22M	77.8	95.9	95.3	99.1
Ours	5.9M	75.8	94.1	92.5	99.1

4.4 Pre-training and fine-tuning

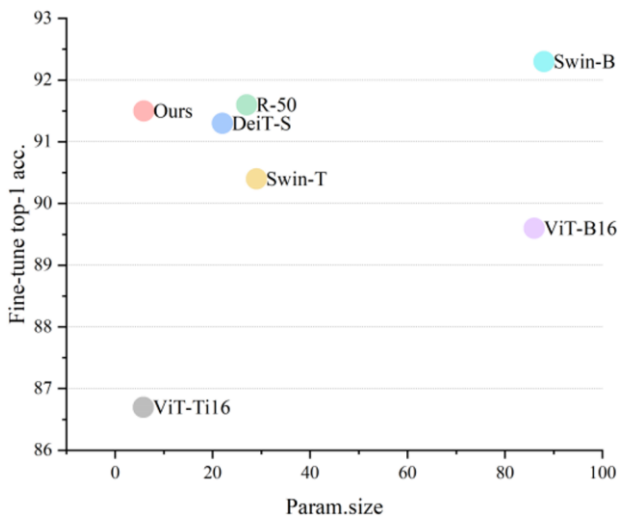


Figure 6. Top-1 accuracy achieved on ImageNet-1k

The methodology for pre-training and fine-tuning aligns with the practices established in literature [13], with adaptations made to accommodate the specific hardware setup used. Pre-training was conducted using the Adam optimizer [20], with β_1 set at 0.9 and β_2 at 0.999. Batch sizes varied in accordance with GPU memory constraints, and a cosine learning rate schedule was employed, incorporating a 10,000-step linear warmup. Random horizontal flipping was used for image preprocessing. On the ImageNet-1k dataset, models were trained over 300 epochs. Fine-tuning involved the use of SGD with a momentum of 0.9, exploring various learning rates and training durations for each dataset.

4.5 Fine-tuning result

All pre-trained models were fine-tuned using the curated dataset of drowning and swimming images. Upon stabilization of the accuracy curve, the backbone module of each model was frozen, with focus shifted solely to training the output module, specifically the MLP heads, to enhance performance. Figure 6 presents the results of the pre-trained models post fine-tuning

with the drowning and swimming images dataset. Detailed top-1 accuracy figures for these results are provided in Table 4.

Table 4. Models' accuracy, precision and recall rate after fine-tuning

Model	Param.	Top-1 acc.%	Precision. %	Recall. %
ViT-Ti16	5.8M	86.7	80.3	86.5
ViT-B16	86M	89.6	85.7	90.2
R-50	27M	91.6	94.1	90.7
Swin-T	29M	90.4	93.5	91.2
Swin-B	88M	92.3	95.5	92.8
DeiT-S	22M	91.3	89.4	92.2
Ours	5.9M	91.5	92.8	90.9

A notable observation is the optimal balance between accuracy and cost achieved by the proposed model. Analysis of the results reveals intriguing insights. The largest model, Swin-B, excels across various metrics, including top-1 accuracy, precision, and recall rates. This performance aligns with the design intent of Swin Transformer, which integrates the strengths of both ViT and CNN models [21]. However, the ViT-B16 and ViT-T16 models display relatively lower performance, particularly in terms of precision. This is hypothesized to stem from the original ViT models' limited local information processing capabilities, adversely affecting their precision. The recall rates of these models partially corroborate this hypothesis.

4.6 MWSDD testing result

Due to resource constraints, including the unavailability of a swimming pool and limited budget for hardware, the effectiveness of the MWSDD system was simulated using two identical cameras. These cameras were positioned to capture video footage in a swimming pool, with their angles set to create a 50% overlap in the filming area, as described in Section 3. This simulation ran for two hours daily over a span of 20 days. During this period, only seven drowning incidents were recorded. These videos were processed and used to fine-tune the proposed model, resulting in an increase in accuracy to 93.2%.

The limited improvement in accuracy through the MWSDD method was anticipated. Two possible explanations are proposed: first, the scarcity of data for training the model; second, the majority of drowning images in the fine-tuning dataset were already correctly identified, suggesting that the primary barrier to accuracy enhancement lies in other aspects of the dataset. A review of the dataset confirmed that 361 out of 387 drowning images were accurately identified.

An interesting observation during the experiment was the necessity for models associated with each camera to begin with the same well-pre-trained parameters but subsequently undergo fine-tuning with data specific to each camera. This is attributed to the unique video data each camera captures; despite a 50% overlap in the filming area, each camera still records 50% unique footage.

In terms of further research and model enhancement, three avenues are identified: (i) conducting additional experiments to ascertain the optimal number of plain CNN layers and pooling layers; (ii) incorporating methodologies from other research, such as distillation or self-supervised training; (iii) refining hyper-parameters for improved performance.

5. CONCLUSION

This study has contributed to the field of drowning detection by developing a hybrid ViT model and introducing the MWSDD method to enhance system accuracy. The research addressed two primary concerns: the balance between cost and efficiency of the model, and the provision of an effective drowning detection system.

Two key innovations were presented. First, a simplified, compact yet efficient model was developed, achieving a 91.5% accuracy on a specialized swimming and drowning dataset, compared to the 89.6% accuracy of the ViT-B16 model. This was accomplished by integrating plain CNNs and pooling layers to replace the patch projection layers of the original ViT, significantly reducing the model's parameter size. Second, the MWSDD method was proposed, leveraging an overlapping field of view between adjacent cameras to improve accuracy.

The applications of the MWSDD technique and the hybrid ViT model are anticipated to be particularly beneficial for small and medium-sized swimming pools, local government facilities, and private pool owners. These entities, seeking cost-effective yet reliable drowning detection solutions, may find these innovations especially relevant.

In summary, this research not only addresses the urgent need for improved drowning detection but also offers practical solutions that balance performance with resource constraints. The advancements made here open avenues for further exploration in enhancing safety measures in aquatic environments.

ACKNOWLEDGMENT

This work is supported by the Industry-University Cooperative Education Project of Ministry of Education of China.

REFERENCE

- [1] World Health Organization. (2017). Preventing drowning: An implementation guide. World Health Organization.
- [2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. <https://doi.org/10.48550/arXiv.2010.11929>
- [3] Shehata, A.M., Mohamed, E.M., Salem, K. L., Mohamed, A.M., Salam, M.A., Gamil, M.M. (2021). A survey of drowning detection techniques. In 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, pp. 286-290. <https://doi.org/10.1109/MIUCC52538.2021.9447677>
- [4] Eng, Toh, Kam, Wang, Yau. (2003). An automatic drowning detection surveillance system for challenging outdoor pool environments. In Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, pp. 532-539. <https://doi.org/10.1109/ICCV.2003.1238393>
- [5] Reddy, S.P.K., Harikiran, J. (2022). Cast Shadow Angle Detection in morphological aerial images using faster R-CNN. Traitement du Signal, 39(4): 1313-1321. <https://doi.org/10.18280/ts.390424>

- [6] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. <https://doi.org/10.48550/arXiv.2004.10934>
- [7] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA.
- [8] Chen, X., Xie, S., He, K. (2021). An empirical study of training self-supervised vision transformers. arXiv:2104.02057. <https://doi.org/10.48550/arXiv.2104.02057>
- [9] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M. (2021). Levit: A vision transformer in convnet's clothing for faster inference. arXiv:2104.01136. <https://doi.org/10.48550/arXiv.2104.01136>
- [10] Shatnawi, M., Albreiki, F., Alkhoori, A., Alhebshi, M. (2023). Deep learning and vision-based early drowning detection. *Information*, 14(1): 52. <https://doi.org/10.3390/info14010052>
- [11] Vestnikov, R., Stepanov, D., Bakhshiev, A. (2023). Development of neural network algorithms for early detection of drowning in swimming pools. In 2023 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Sochi, Russian Federation, pp. 820-824. <https://doi.org/10.1109/ICIEAM57311.2023.10139153>
- [12] He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385. <https://doi.org/10.48550/arXiv.1512.03385>
- [13] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L. (2021). How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270. <https://doi.org/10.48550/arXiv.2106.10270>
- [14] Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V. (2012). Cats and dogs. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, pp. 3498-3505. <https://doi.org/10.1109/CVPR.2012.6248092>
- [15] Krizhevsky, A., Hinton, G. (2009). Learning multiple layers of features from tiny images. University of Toronto. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>
- [16] Nilsback, M.E., Zisserman, A. (2008). Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, pp. 722-729. <https://doi.org/10.1109/ICVGIP.2008.47>
- [17] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [18] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 10012-10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [19] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, pp. 10347-10357.
- [20] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. <https://doi.org/10.48550/arXiv.1412.6980>
- [21] Sun, C., Shrivastava, A., Singh, S., Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 843-852. <https://doi.org/10.1109/ICCV.2017.97>