RetinaGate: A Gated Feature Pyramid Network for Improved Object Detection with SE-based Attention

Mahtab Jamali[1,*,†], Paul Davidsson[1,†], Reza Khoshkangini[1,†], Martin Georg Ljungqvist[2,†] and Radu-Casian Mihailescu[1,†]

[1]Department of Computer Science and Media Technology, Sustainable Digitalisation Research Centre, Malmö University, Malmö, Sweden

[2]Axis Communications AB, Lund, Sweden

Abstract

Object detection is a critical task in computer vision with wide-ranging applications, from autonomous driving to surveillance systems. Despite notable progress, challenges such as detecting small objects, managing occlusions, and effectively integrating multiscale features persist. We propose RetinaGate, a novel object detection architecture that introduces a Gated Feature Pyramid Network (G-FPN) to adaptively fuse multi-scale features, enhanced by Squeeze-and-Excitation-based channel attention for improved accuracy. As a plug-and-play module, G-FPN can be seamlessly integrated into existing detection models to enhance their accuracy. These enhancements strengthen the model's capacity to capture fine-grained details and leverage contextual information more effectively. Experimental results on three benchmark datasets demonstrate that RetinaGate outperforms the baseline RetinaNet in terms of detection accuracy, particularly in challenging detection scenarios such as underwater.

Keywords

Object Detection, RetinaNet, FPN, Gated Fusion, RetinaGate, SEBlock

1. Introduction

Object detection has become a cornerstone in the field of computer vision, with wide-ranging applications that include autonomous driving, medical diagnostics [1], and real-time video analysis [2]. As an essential component of intelligent systems, object detection aims to locate and classify objects within an image, making it crucial for tasks requiring both precision and computational efficiency [3, 4].

While deep learning detectors such as RetinaNet [5], Faster R-CNN [6], and YOLO [7] have achieved remarkable progress, some challenges persist. Small object detection [8] remains a significant hurdle due to insufficient feature resolution at higher pyramid levels. Other challenges include occlusion, where objects are partially hidden from view, and cluttered backgrounds, which can lead to false positives or missed detections [9]. Furthermore, the semantic gap between low-level and high-level features can hinder precise localization and classification, especially in complex environments [10, 11]. These limitations highlight the need for more sophisticated backbone architectures and robust feature fusion mechanisms to improve detection accuracy across diverse scenarios.

RetinaNet [12], a one-stage object detector known for its efficiency and Focal Loss, provides a robust baseline for addressing common detection challenges. However, its default architecture can be further enhanced to improve performance in complex scenarios, such as detecting small or occluded objects. One limitation of the standard ResNet-50 backbone is its inability to adaptively focus on the most informative feature channels, which can reduce its effectiveness in cluttered or context-rich scenes. In addition, the standard Feature Pyramid Network (FPN) processes each pyramid level independently, without explicitly fusing cross-level information. This limits its ability to fully exploit the complementary strengths of multi-scale features.

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.

*Corresponding author.

†These authors contributed equally.

mahtab.jamali@mau.se (M. Jamali[1,*,†]); paul.davidsson@mau.se (P. Davidsson[1,†]); reza.khoshkangini@mau.se (R. Khoshkangini[1,†]); martin.ljungqvist@axis.com (M. G. Ljungqvist[2,†]); radu.c.mihailescu@mau.se (R. Mihailescu[1,†])

To address these limitations, we propose a novel enhancement to RetinaNet, titled RetinaGate by incorporating Squeeze-and-Excitation (SE) blocks [13] and a novel FPN, titled G-FPN (Gated Fusion FPN). SE block, integrated into the ResNet-50 backbone, improve channel-wise attention, enabling the model to prioritize the most informative features. The Gated Fusion module, applied after the Feature Pyramid Network (FPN), enhances the fusion of multiscale features, ensuring robust performance across diverse object sizes and challenging conditions. These modifications specifically target the weaknesses in handling small objects, occlusions, and the integration of multiscale features, which are critical for achieving higher detection accuracy.

This paper is structured as follows. In Section 2, we discuss related works, focusing on advancements in backbone architectures, feature fusion techniques, and one-stage detectors. Section 3 outlines the methodology behind our proposed enhancements, detailing the integration of SE blocks and Gated Fusion. Section 4 presents the datasets used for evaluation, and Section 5 reports the experimental results, demonstrating the superiority of RetinaGate over the baseline RetinaNet. Finally, Section 6 concludes with future research directions.

2. Related works

The field of object detection has witnessed substantial progress with the development of various architectures and techniques. Among them, RetinaNet has stood out as a significant contribution, offering a balance between accuracy and computational efficiency. However, several studies have identified limitations in RetinaNet and proposed enhancements to address them:

RetinaNet and Multiscale Detection: Lin et al. introduced RetinaNet with Focal Loss to mitigate the impact of class imbalance in object detection [5]. Despite its success, challenges such as small object detection and effective multiscale feature integration remain. For instance, the work by Kong et al. introduced the Deep Feature Pyramid Network (DFPN) [14], which augments FPNs with enhanced connectivity to improve multiscale detection, particularly for small objects. Similarly, Libra R-CNN [15] addressed multiscale imbalance by introducing balanced feature pyramid integration. The NAS-FPN [16] utilized neural architecture search to optimize feature pyramid designs, achieving state-of-the-art performance. However, these solutions often introduce significant computational complexity. BiFPN, proposed in EfficientDet [17], enhanced multiscale detection by employing lightweight, bidirectional feature fusion. While effective, BiFPN requires fine-tuned hyperparameters and is not tailored for one-stage detectors like RetinaNet.

Feature Enhancement Mechanisms: Researchers have proposed various mechanisms to enhance feature representation. Hu et al. introduced Squeeze-and-Excitation Networks to recalibrate channel-wise feature responses dynamically. These networks have been integrated into different architectures to improve attention mechanisms. For example, SENet [18] was successfully applied to image classification tasks, and Zhang et al. (2020) extended it to Faster R-CNN for improving object detection. Similarly, Woo et al. proposed Convolutional Block Attention Module (CBAM) [19], which combines channel and spatial attention for enhanced feature extraction. CBAM has been incorporated into architectures like YOLOv4, demonstrating improvements in feature selectivity. More recently, Efficient Attention Networks (EANet) [20] introduced lightweight attention mechanisms for real-time object detection, which significantly reduced computational overhead. However, these methods have primarily focused on classification tasks or two-stage detectors, with limited exploration in one-stage models like RetinaNet.

Contextual and Multiscale Fusion: Feature fusion is another area of focus for improving object detection [21]. Works such as PANet [22], and NAS-FPN [23] emphasize enhancing information flow across scales. PANet introduced bottom-up path augmentation to complement FPN's top-down feature flow, improving multiscale detection capabilities. More recently, Auto-FPN [24] employed neural architecture search to automatically design efficient feature fusion paths, addressing multiscale detection while maintaining computational efficiency. Additionally, Dynamic FPN [25] integrated adaptive mechanisms to dynamically adjust the contributions of feature levels based on the input image characteristics, further enhancing context-aware fusion. Although effective, these approaches often

involve high computational costs, making them less suitable for real-time applications. Our approach incorporates a Gated Fusion module, which selectively integrates multiscale features while maintaining efficiency, addressing both contextual relevance and multiscale challenges.

Enhancements in FPN Design: Enhancements to the original FPN architecture have focused on improving information flow and balancing feature contributions. Libra R-CNN [15] introduced a balanced semantic path to reduce feature-level imbalance, significantly improving object detection across scales. NAS-FPN [23] used neural architecture search to automate FPN design, resulting in high-performing but computationally expensive structures. BiFPN, proposed in EfficientDet [17], employed bidirectional fusion to refine multiscale feature integration while reducing computational cost. Additionally, works like Path Aggregation Network (PANet) [22] extended FPN with bottom-up paths, enabling improved feature reuse for instance segmentation and detection tasks. Recently, Dynamic FPN [25] adapted FPN contributions dynamically based on input image requirements, addressing both efficiency and adaptability. While these approaches provide valuable insights, many require extensive computational resources or are highly domain-specific, limiting their generalizability. Our work adopts a simpler, yet effective Gated Fusion strategy, ensuring scalability and efficiency for diverse detection tasks.

Related Enhancements in One-Stage Detectors: One-stage object detectors such as YOLO[26], SSD [27], and RetinaNet have been the subject of extensive research and development. SSD (Single Shot MultiBox Detector) introduced a novel approach to predict object locations and class scores directly from feature maps, leveraging multiple feature scales for detecting objects of various sizes. However, its fixed anchor configurations posed challenges for small object detection. YOLOv3 and its successors, YOLOv4 [28] and YOLOv5 [29], addressed these limitations by employing improved feature extraction backbones such as CSPNet and introducing techniques like mosaic augmentation to enhance training data diversity. YOLOv7 [30] and YOLOv8 further explored decoupled head architectures, lightweight attention modules, and optimized training pipelines to improve accuracy and efficiency. Similarly, FCOS (Fully Convolutional One-Stage Object Detection) removed the need for anchor boxes altogether, relying on a center-ness score to predict object locations directly, thus simplifying the pipeline while maintaining competitive performance. Despite these advances, integrating robust feature attention and fusion mechanisms, as proposed in our work, remains a critical gap for improving small and occluded object detection in one-stage detectors.

Our work differentiates itself by integrating Squeeze-and-Excitation blocks with Gated Fusion directly into RetinaNet's architecture. By addressing the limitations of both the backbone and FPN, our approach provides a comprehensive solution for object detection and multiscale feature integration without incurring significant computational overhead. Additionally, our method uniquely combines adaptive feature prioritization and gated feature fusion, filling the gap between lightweight design and robust feature representation.

3. Approach

3.1. Overview of the Proposed Approach

The proposed approach is illustrated in Figure 1, comprising four main components: (a) ResNet Backbone, (b) SE Blocks, (c) Feature Pyramid Network (FPN), (d) G-FPN. This architecture combines SEblock and the novel G-FPN (Gated Fusion FPN) which contains a gated fusion module to address challenges such as small object detection, occlusions, and domain-specific variations, resulting in enhanced detection accuracy and robustness. Unlike a standard FPN that directly passes feature maps to the classification and regression heads without additional refinement, G-FPN integrates the Gated Fusion module to generate an enriched feature map. This additional feature map enhances the multiscale feature representation, improving detection accuracy by enabling better contextual understanding and feature refinement. The proposed model consists of the following components:

(a) ResNet Backbone: The ResNet-50 backbone extracts hierarchical feature maps from the input image, capturing both low-level and high-level representations.

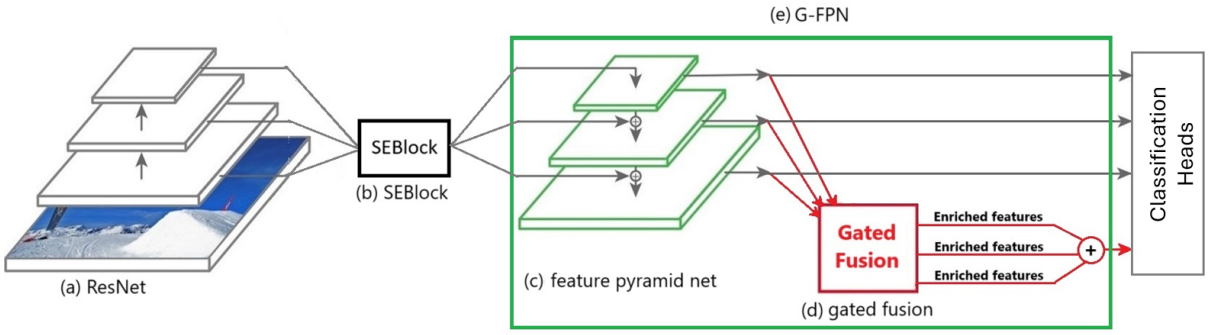

Figure 1: The architecture of the proposed approach that contains four main components: (a) ResNet Backbone, (b) SE Blocks, (c) Feature Pyramid Network (FPN), (d) G-FPN.

(b) SE Blocks: Squeeze-and-Excitation (SE) block is incorporated into the ResNet-50 backbone after each major layer group (layer1, layer2, layer3, and layer4). This block enhances the model's capability to recalibrate channel-wise feature responses adaptively by modeling dependencies between channels. By prioritizing informative features and suppressing less relevant ones, SE blocks improve the robustness of feature representations.

The primary reason for placing SE blocks in ResNet-50 is to enhance hierarchical feature learning across different layers:

- Early-Layer Enhancement: SE blocks in lower layers focus on improving edge and texture details, critical for small object detection.
- Mid-Layer Refinement: At intermediate layers, they refine semantic feature representation for medium-sized objects.
- Deep-Layer Contextualization: In the final layer group, SE blocks emphasize high-level semantic features, which are essential for addressing occlusions and complex object shapes.

This block integration ensures that features at all scales are adaptively weighted, contributing to improved multiscale detection performance.

(c) Feature Pyramid Network (FPN): The FPN aggregates multiscale features from the ResNet backbone, enabling robust detection of objects at varying scales.

(d) G-FPN (Gated Fusion FPN): The original FPN aggregates multiscale features without any dynamic weighting, treating all scales equally. In contrast, G-FPN introduces:

- Dynamic Feature Prioritization: Ensures relevant scales contribute more significantly.
- Enhanced Feature Representation: Combines fused features with original multiscale outputs, providing richer context.
- Plug-and-Play Flexibility: Can be integrated into various detection architectures without significant modification.

By dynamically weighting the contributions of each scale, G-FPN ensures that the most relevant features are prioritized, improving the model's ability to handle objects of varying sizes and complexities.

The structure of G-FPN is shown in Figure 2. The Gated Fusion Module is designed to enhance the integration of multiscale feature maps, enabling adaptive fusion based on feature relevance. Unlike the standard FPN, which aggregates features in a static manner, gated fusion module incorporates gating mechanisms to modulate the contributions of each scale dynamically. This ensures a more context-aware and robust feature representation. The Gated Fusion module integrates feature maps from multiple levels of the Feature Pyramid Network (FPN) and produces an enriched feature map. The gated fusion is computed as:

The Gated Fusion Module is designed to selectively integrate feature maps from multiple levels of the Feature Pyramid Network (FPN). It achieves this by employing a gating mechanism that dynamically adjusts the contributions of individual feature maps to the final fused representation.

Mathematical Formulation Let $F_1, F_2, \ldots, F_n$ represent the feature maps from n levels of the FPN, where $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, and C_i, H_i, and W_i are the channel, height, and width dimensions of the feature map at level i. The Gated Fusion Module combines these feature maps as follows:

1. **Spatial Alignment:** Each feature map is resized to a common spatial resolution, denoted as (H_r, W_r), which corresponds to the resolution of a reference feature map (e.g., the first feature map, F_1):
$$\hat{F}_i = \text{Interpolate}(F_i, \text{size} = (H_r, W_r), \text{mode='nearest'}),$$
where $\hat{F}_i$ is the resized feature map at level i.
2. **Gating Mechanism:** For each resized feature map $\hat{F}_i$, a gating mechanism is applied to compute the importance weights. The gating function is defined as:
$$G(\hat{F}_i) = \sigma\left(W_2 * \text{ReLU}\left(W_1 * \hat{F}_i\right)\right),$$
where:
   - $W_1 \in \mathbb{R}^{C \times (C/r) \times 1 \times 1}$ and $W_2 \in \mathbb{R}^{(C/r) \times C \times 1 \times 1}$ are learnable weight tensors.
   - r is the reduction ratio, which controls the dimensionality reduction in the gating mechanism.
   - $*$ denotes convolution, and ReLU$(\cdot)$ is the Rectified Linear Unit activation function.
   - $\sigma(\cdot)$ represents the sigmoid activation function, which scales the importance weights between 0 and 1.
3. **Feature Weighting:** The gated feature map is obtained by element-wise multiplication of the gating weights and the resized feature map:
$$\hat{F}_i^{\text{gated}} = G(\hat{F}_i) \odot \hat{F}_i,$$
where $\odot$ denotes element-wise multiplication.
4. **Feature Fusion:** The final fused feature map is computed by summing the gated feature maps from all levels:
$$F_{\text{fused}} = \sum_{i=1}^{n} \hat{F}_i^{\text{gated}}.$$

The gating mechanism adaptively learns the importance of features at each level of the FPN, ensuring that only the most relevant features contribute to the final fused representation. The interpolation step aligns the spatial dimensions of all feature maps, enabling effective fusion across scales. The reduction ratio r controls the complexity of the gating mechanism, allowing for efficient computation.

Advantages

- Enables selective emphasis on important features from different levels of the FPN.
- Facilitates multi-scale feature integration, enhancing the network's ability to capture both fine and coarse details.
- Reduces the impact of redundant or irrelevant features, improving the overall performance of the object detection model.

4. Datasets

To evaluate the performance and generalization capability of our proposed model, we conducted experiments on three datasets: Pascal VOC 2007, Pascal VOC 2012, and the Aqua dataset. These datasets encompass a range of object categories and challenging conditions, allowing us to demonstrate the versatility and robustness of our enhancements.

1. Pascal VOC 2007

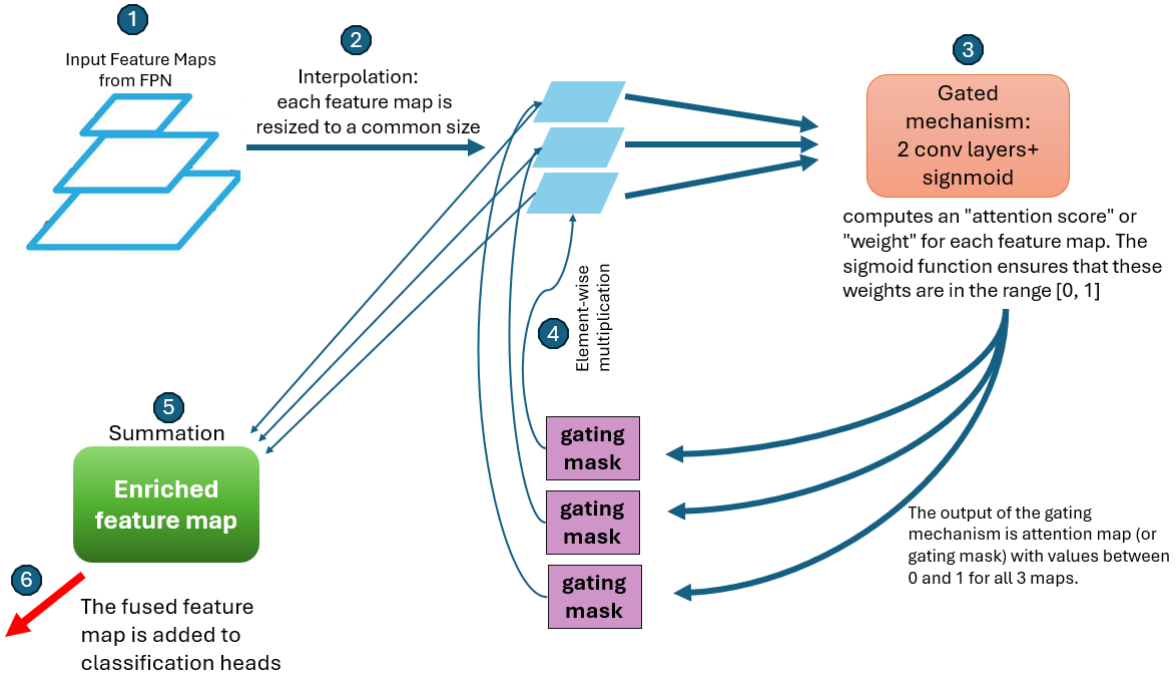

Figure 2: G-FPN architecture

Pascal VOC 2007 [31] consists of 5,000 training images and 4,900 testing images, covering 20 object categories. We performed an ablation study using subsets of Pascal VOC 2007 to analyze the effectiveness of our modifications. Initially, we tested the model with 100 images from three classes (person, car, bus) enabling a focused evaluation of the model's improvements in a simplified setting. Subsequently, we increased the dataset to 1,000 images covering four classes (person, car, bus, motorbike) to examine the scalability and consistency of the enhancements. Finally, the model was evaluated on the complete Pascal VOC 2007 dataset to assess its generalization capability across diverse object classes and a larger number of images.

2. Pascal VOC 2012

Pascal VOC 2012 [32] consists of 13,690 training images and 3,422 testing images, providing a more comprehensive dataset compared to Pascal VOC 2007. This dataset includes additional images and variations in image conditions, allowing us to validate the model's generalization ability across different distributions. Testing on Pascal VOC 2012 ensures the robustness of our approach in handling diverse object classes and environmental variations.

3. Aqua Dataset

The Aqua dataset contains 575 training images and 63 testing images, specifically designed for underwater object detection. This dataset presents unique challenges, such as blurred objects, low visibility, and occlusions caused by underwater conditions. These factors often complicate the detection of marine life, such as fish, which are not only camouflaged but also exhibit irregular shapes and movements. By applying our model to this dataset, we demonstrate its adaptability and capability to handle complex environments outside the standard datasets used for object detection.

5. Results

We conducted an ablation study using the Pascal VOC 2007 dataset, progressively analyzing the impact of SEblock and G-FPN on the baseline RetinaNet model. The study included testing with subsets of Pascal VOC 2007 (100 images and 1,000 images) and the complete Pascal VOC 2007 dataset to understand the contribution of each module. Additionally, the complete approach (Model 4) was evaluated on Pascal VOC 2012 and the Aqua dataset to assess its generalization across different domains and challenging

scenarios. For all three complete datasets, we trained the proposed model five times and calculated the standard deviation to demonstrate the stability of the results.

The Table 1 presents the mean Average Precision (mAP) results for different configurations:

Model	300 images	1000 images	Pascal 2007	Pascal 2012	Aqua dataset
Model 1: Original RetinaNet	41.24	47.09	56.44	53.32	61.64
Model 2: Adding SEBlock	42.52	48.97	57.54	-	-
Model 3: Adding Gated Fusion	43.59	46.35	56.83	-	-
Model 4: Adding SEBlock and Gated Fusion	**45.73**	**48.54**	**57.86**	**54.74**	**64.18**

Table 1
Comparing mAP of the proposed model with the original RetinaNet on various datasets.

Key Insights:

- Pascal VOC 2007 Analysis: For 300 images (3 classes), adding SE blocks alone improved mAP by 1.28%, while G-FPN alone contributed a 2.35% increase. The addition of SE blocks to the ResNet backbone and G-FPN improved detection performance incrementally. The proposed model (Model 4) achieved the highest mAP, showcasing the effectiveness of combining SEblock and G-FPN.
- Pascal VOC 2012 and Aqua Dataset: Model 4 was further tested on Pascal VOC 2012 and the Aqua dataset to evaluate generalization across different domains. The proposed model outperformed the baseline RetinaNet and other configurations, achieving higher mAP in all scenarios. This highlights its robustness in handling diverse object classes and challenging conditions, such as underwater environments where objects may appear blurred or obstructed.

The following table presents the comparison of our proposed approach with are methods across the Pascal VOC 2007, Pascal VOC 2012, and Aquarium datasets.

Dataset	Method	mAP
Pascal VOC 2007	RetinaGate (ours)	**57.86**
	FemtoDet [33]	46.31
	Deformable Parts Model [34]	45.20
	TinyissimoYOLO-v8 [35]	42.30
Pascal VOC 2012	RetinaGate (ours)	**54.74**
	CenterNet [36]	47.00
	DETR [37]	54.30
Aquarium Dataset	RetinaGate (ours)	**64.18**
	SCL [38]	0.349
	SCAN [39]	0.545
	SIGMA [40]	0.636
	YOLOv5 [41]	0.516

Table 2
Comparison of mAP values for different datasets and methods.

6. Conclusion

In this paper, we presented RetinaGate, an enhanced RetinaNet-based object detection model incorporating Squeeze-and-Excitation (SE) blocks and a novel FPN, Gated Fusion FPN (G-FPN). By integrating SE blocks into the ResNet-50 backbone and introducing G-FPN for adaptive multiscale feature fusion, our approach effectively addressed challenges such as small object detection, occlusions, and complex feature integration.

Experimental results across Pascal VOC 2007, Pascal VOC 2012, and the Aquarium dataset demonstrated the superiority of the proposed model compared to baseline RetinaNet and several state-of-the-art methods. Our results highlight the strength of the G-FPN as a plug-and-play module that can be integrated into other architectures to improve detection performance, particularly in scenarios involving challenging domains such as underwater environments where objects are often blurred or occluded. This flexibility and the observed performance gains underline the potential of our proposed enhancements for broader applications in object detection tasks. Future research will focus on further evaluating the generalizability of the G-FPN across more diverse datasets and exploring its integration into other backbone architectures to fully leverage its capabilities.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for some sections to check grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Madhavan, Object detection human activity recognition for improved patient mobility and caregiver ergonomics (2025).

[2] M. Jamali, P. Davidsson, R. Khoshkangini, M. G. Ljungqvist, R.-C. Mihailescu, Context in object detection: a systematic literature review, Artificial Intelligence Review 58 (2025) 1–89.

[3] M. Jamali, P. Davidsson, R. Khoshkangini, M. G. Ljungqvist, R.-C. Mihailescu, Specialized indoor and outdoor scene-specific object detection models, in: Sixteenth International Conference on Machine Vision (ICMV 2023), volume 13072, SPIE, 2024, pp. 201–210.

[4] M. Jamali, P. Davidsson, R. Khoshkangini, R.-C. Mihailescu, E. Sexton, V. Johannesson, J. Tillström, Video-audio multimodal fall detection method, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 2024, pp. 62–75.

[5] T.-Y. Ross, G. Dollár, Focal loss for dense object detection, in: proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2980–2988.

[6] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE transactions on pattern analysis and machine intelligence 39 (2016) 1137–1149.

[7] P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A review of yolo algorithm developments, Procedia computer science 199 (2022) 1066–1073.

[8] Y. Liu, P. Sun, N. Wergeles, Y. Shang, A survey and performance evaluation of deep learning methods for small object detection, Expert Systems with Applications 172 (2021) 114602.

[9] Z. Li, Y. Dong, L. Shen, Y. Liu, Y. Pei, H. Yang, L. Zheng, J. Ma, Development and challenges of object detection: A survey, Neurocomputing 598 (2024) 128102.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[11] M. Tajgardan, A. Shiranzaei, M. Jamali, R. Khoshkangini, M. Rabbani, Advanced stock market prediction using unsupervised federated learning techniques, in: 2025 29th International Computer Conference, Computer Society of Iran (CSICC), IEEE, 2025, pp. 1–6.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.

[13] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[14] F. Hou, Q. Gao, Y. Song, Z. Wang, Z. Bai, Y. Yang, Z. Tian, Deep feature pyramid network for eeg emotion recognition, Measurement 201 (2022) 111724.
[15] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra r-cnn: Towards balanced learning for object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 821–830.
[16] G. Ghiasi, T.-Y. Lin, Q. V. Le, Nas-fpn: Learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7036–7045.
[17] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.
[18] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141. doi:10.1109/CVPR.2018.00745.
[19] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19. doi:10.1007/978-3-030-01234-2_1.
[20] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Efficient attention network: Accelerate attention by searching where to plug, arXiv preprint arXiv:2206.01659 (2022).
[21] R. Khoshkangini, M. Tajgardan, M. Jamali, M. G. Ljungqvist, R.-C. Mihailescu, P. Davidsson, Hierarchical transfer multi-task learning approach for scene classification, in: International Conference on Pattern Recognition, Springer, 2024, pp. 231–248.
[22] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8759–8768. doi:10.1109/CVPR.2018.00913.
[23] G. Ghiasi, T.-Y. Lin, Q. V. Le, Nas-fpn: Learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7036–7045. doi:10.1109/CVPR.2019.00721.
[24] Q. Wang, T. Yang, J. Zhang, Z. Li, Y. Chen, J. Wang, J. Sun, Auto-fpn: Automatic network architecture adaptation for object detection beyond classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6649–6658. doi:10.1109/CVPR42600.2020.00668.
[25] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Dynamic feature pyramid networks for object detection, arXiv preprint arXiv:2012.00779 (2021).
[26] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).
[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 21–37. doi:10.1007/978-3-319-46448-0_2.
[28] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020).
[29] G. Jocher, A. Chaurasia, J. Qiu, YOLO by Ultralytics, 2020. URL: https://github.com/ultralytics/yolov5.
[30] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, arXiv (????).
[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.
[32] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, International journal of computer vision 111 (2015) 98–136.
[33] P. Tu, X. Xie, G. Ai, Y. Li, Y. Huang, Y. Zheng, Femtodet: An object detection baseline for energy versus performance tradeoffs, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13318–13327.

[34] R. Girshick, F. Iandola, T. Darrell, J. Malik, Deformable part models are convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 437–446.
[35] J. Moosmann, P. Bonazzi, Y. Li, S. Bian, P. Mayer, L. Benini, M. Magno, Ultra-efficient on-device object detection on ai-integrated smart glasses with tinyissimoyolo, arXiv preprint arXiv:2311.01057 (2023).
[36] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6569–6578.
[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.
[38] A. Löchner, Semantic clustering by adopting nearest neighbor (scan), in: Der andere Sport: Esports zwischen gesellschaftlichem Strukturwandel und Marketingstrategie, Springer, 2025, pp. 365–388.
[39] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, arXiv preprint arXiv:1808.06670 (2018).
[40] W. Li, X. Liu, Y. Yuan, Sigma: Semantic-complete graph matching for domain adaptive object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5291–5300.
[41] G. Jocher, Ultralytics yolov5, 2020. URL: https://github.com/ultralytics/yolov5. doi:10.5281/zenodo.3908559.