

A Deep Learning Approach To Human Fall Detection With YOLOv12 Versi Nano

Floresce Sumarauw¹, Febriyanti Ludja², Robby Moody Lintong³

¹Master Program of Informatics, Postgraduate Program, Sam Ratulangi University Manado, Kampus Bahu St., 95115, Indonesia

² Master Program of Informatics, Postgraduate Program, Sam Ratulangi University Manado, Kampus Bahu St., 95115, Indonesia

³ Master Program of Informatics, Postgraduate Program, Sam Ratulangi University Manado, Kampus Bahu St., 95115, Indonesia

Corresponding e-mail: florescesumarauw026@student.unsrat.ac.id

Abstract—The growing elderly population faces a heightened risk of falls. Current monitoring solutions, such as in-person observation by caregivers or wearable sensors, suffer from several drawbacks. These include low accuracy, user discomfort, and a constant need for human oversight. To overcome these issues, automated approaches using Convolutional Neural Networks (CNNs), especially the YOLO algorithm, have been applied for fast and real-time fall detection. In this study, the original YOLOv12-nano model, which is known for its lightweight architecture, low latency, and efficiency on edge devices, was implemented on the Le2i dataset. The model achieved an mAP50 of 0.866 and an mAP50-95 of 0.555, with only 2.52 M parameters and 6.0 GFLOPs. These provide a strong foundation for further improvement in elderly fall detection systems.

Keywords—Fall Detection, Convolutional Neural Network, YOLOv12, Deep Learning

I. INTRODUCTION

The global elderly population is increasing rapidly, making older adults the fastest-growing age group worldwide. As of October 2022, one in ten people worldwide was aged 65 years or older [1], and this number is projected to reach approximately 1.5 billion by 2050 [2]. Aging is associated with declines in physical strength, cognitive function, and motor coordination, which significantly increase the risk of falls, a major cause of injury and mortality among older adults [3]. According to the World Health Organization (WHO), approximately 37.3 million falls requiring medical attention occur annually, resulting in around 650,000 deaths [4]. Timely fall detection can reduce mortality risk and improve recovery outcomes, highlighting the importance of accurate real-time fall detection systems [5].

Conventional fall monitoring methods, including caregiver supervision and wearable sensors, have several limitations. Manual monitoring is labor-intensive and prone to human error, while wearable devices may cause discomfort and provide limited reliability [6]. Consequently, computer vision has emerged as a promising alternative for contact-free and automated monitoring. Object detection enables the localization of people in images and video, making it well suited for real-time fall detection applications [7], [8].

Convolutional Neural Networks (CNNs) have become the dominant approach for object detection because of their ability to learn discriminative visual features [9], [10]. Among CNN-based detectors, YOLO (You Only Look Once) is widely adopted because of its high detection accuracy and real-time inference capability [11], [12]. The latest YOLOv12-nano [13] employs a lightweight architecture with low computational complexity, making it suitable for deployment on edge devices.

In this study, YOLOv12-nano [13] is applied to fall detection using the Le2i dataset [14]. The model achieves an mAP50 of 0.866 and an mAP50:95 of 0.555 with only 2.52 million parameters and 6.0 GFLOPs, demonstrating a favorable balance between detection accuracy and computational efficiency.

II. METHOD

This study proposes the application of YOLOv12-Nano as a solution for real-time human fall detection in complex environments [15-19]. This architecture was selected for its high computational efficiency and ability to maintain optimal detection accuracy, making it suitable for deployment on devices with limited computational resources [20]. As a lightweight variant of the YOLO family, YOLOv12-Nano is structured to achieve a balance between inference speed and detection accuracy. As shown in Figure 1, this architecture comprises three main components: Backbone, Neck, and Head.

A. Backbone

The backbone in YOLOv12 serves as the main component in extracting features from input images. The image data gained will be processed through a series of convolution layers to form a representation in the form of feature maps at various scales. In this version, R-ELAN (Residual Efficient Layer Aggregation Network) serves as the core block of the backbone architecture. R-ELAN aims to retain feature information from the initial layers, enhance gradient propagation efficiency during training, and reduce computational complexity compared to conventional backbones. Through these innovations, YOLOv12 can maintain training stability in a deeper architecture without sacrificing speed or accuracy during inference.

B. Neck

The Neck in YOLOv12 plays a crucial role in integrating information from multiple feature scales generated by the backbone. This process is realized through the combination of the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN), enabling the network to preserve both spatial and semantic representations of objects of varying sizes, from small to large. In addition, YOLOv12 incorporates the Area Attention (A^2) and Flash Attention mechanisms to enhance the model's focus on critical regions, such as small or overlapping objects, thereby improving spatial attention efficiency and reducing potential information loss during feature aggregation. Through this approach, the Neck not only serves as a multi-scale feature fusion module but also emphasizes the most relevant regions to support detection accuracy.

C. Head

The Head in YOLOv12 is responsible for producing the final predictions, including bounding box positions, confidence scores, and object categories. Its design focuses on maintaining inference speed without compromising detection accuracy. Key innovations include the adoption of a decoupled head structure, which separates the classification and bounding box regression processes to prevent optimization conflicts, and the implementation of Convolution-BatchNorm-SiLU (CBS) blocks as fundamental units to ensure stable data flow. Furthermore, YOLOv12 employs optimized loss functions, such as Task-Aligned Assigner and Optimal Transport Assignment, to accelerate convergence and enhance prediction quality, particularly for overlapping objects.

D. Implementation and Setup

To evaluate the effectiveness of the proposed method, a series of experiments was conducted to achieve optimal results while maintaining a balance between computational

efficiency and model accuracy. The system configuration and dataset design were carefully planned to align with the research objectives and effectively support the model training process. Further details regarding the implementation stages are presented in the following section.

- Training and testing Configuration

In this study, training and testing of the YOLOv12-Nano model were carried out based on the configuration summarized in Table 1. The training process was conducted on the Kaggle platform using an NVIDIA Tesla P100 GPU, which provides high computational capacity to support efficient training and inference. The input image resolution was defined to 640×640 pixels to preserve a balance between spatial detail accuracy and computational efficiency. The model was trained for 300 epochs with a batch size of 32, ensuring stable convergence while optimizing GPU memory usage. Optimization was performed using the Stochastic Gradient Descent (SGD) algorithm due to its ability to achieve consistent convergence in object detection tasks. A learning rate of 0.01 was applied to accelerate convergence and minimize the risk of overfitting, thereby enhancing the model's generalization capability.

Model testing, evaluation, and inference were conducted on a local device using a CPU. This approach was applied to assess the model's capability for deployment in real-world scenarios with limited computational resources.

- Dataset

The dataset used in this study is the Le2i Fall Detection dataset, which was obtained from previous research [14]. This dataset comprises 3,010 frames, divided into 2,557 for training, 303 for testing, and 150 for validation, as shown in Table 2.

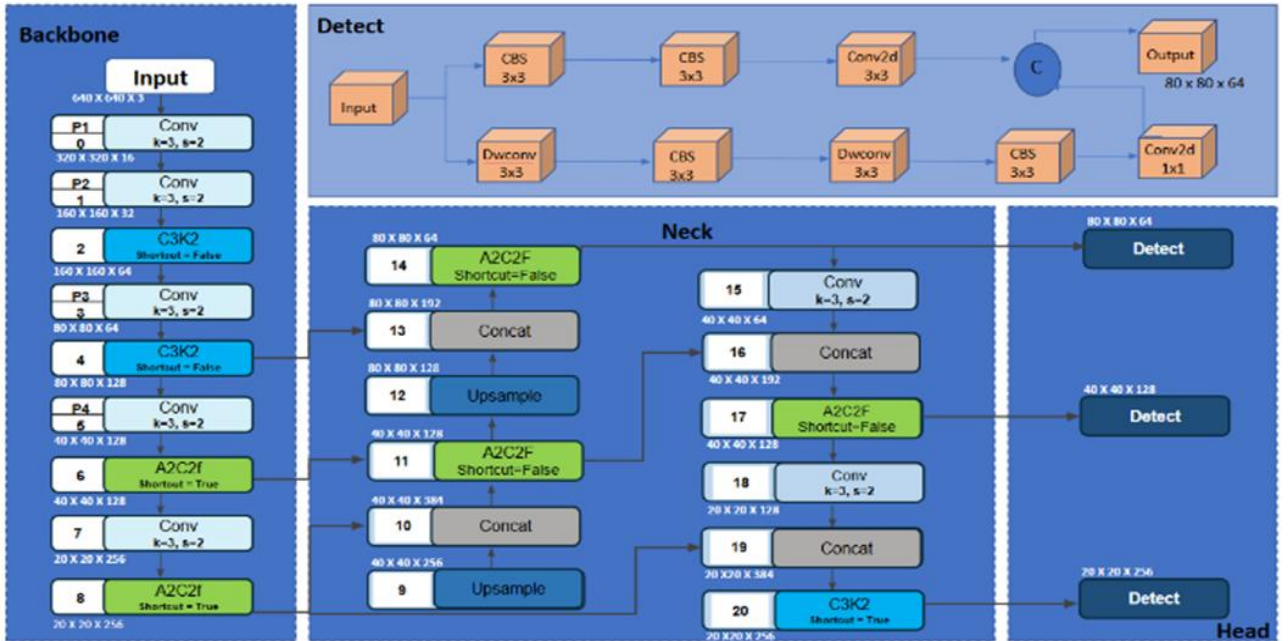


Fig. 1. Architecture of YOLOv12-Nano.



Fig. 2. Sample Images from the Le2i Dataset.

TABLE I. TRAINING AND TESTING CONFIGURATION

<i>Parameters</i>	<i>Setup</i>
Platform/device	Kaggle
GPU	P100
Image Size	640 x 640 pixels
Epochs	300
Batch Size	32
Optimizer	Stochastic Gradient Descent (SGD)
Learning Rate	0,01

TABLE II. DATASET CONFIGURATION

<i>Parameters</i>	<i>Setup</i>
Training Data	2,557 images
Validation Data	303 images
Testing Data	150 images

TABLE III. PERFORMANCE AND EFFICIENCY RESULTS

<i>Model</i>	<i>GFLOPS</i>	<i>Parameter</i>	<i>mAP 50%</i>	<i>mAP 50-95%</i>
YOLOv12-Nano	6.0	2.52	86.6	55.5
YOLO11-Nano	6.4	2.59	88.7	56.0
YOLOv10-Nano	8.4	2.70	85.9	56.5

The data augmentation process was applied exclusively to the training set to enhance the model's generalization capability. Meanwhile, the validation and testing sets remained unchanged to ensure objective performance evaluation. This data-splitting strategy resulted in a balanced, representative dataset, supporting optimal model training and accurate, consistent performance assessment. Each sample in the dataset was annotated with bounding boxes to identify and localize falling subjects. The dataset contains images captured under varying viewpoints and lighting conditions, as well as diverse environmental backgrounds. Furthermore, it poses significant detection challenges, including dynamic subject movements and visual distractions, as illustrated in Figure 2.

Overall, the dataset comprises 3,010 images with approximately 2,557 annotated instances of falling subjects. Each instance represents a fall event within an image, either as a single subject or as part of a group.

III. RESULTS AND DISCUSSION

The performance of the proposed YOLOv12-Nano model was evaluated in terms of detection precision, computational efficiency, robustness under different environmental conditions, and inference efficiency. The evaluation was conducted using a fall detection dataset containing variations in viewpoints, illumination, and background complexity to assess the model's ability to detect fall events under realistic conditions.

A. Evaluation on the Dataset

The detection performance of YOLOv12-Nano was evaluated using mean Average Precision at an IoU threshold of 0.5 (mAP50) and over the IoU range of 0.5–0.95 (mAP50:95). The model achieved an mAP50 of 86.6% and an mAP50:95 of 55.5%, with 2.52 million parameters and 6.0 GFLOPs. For comparison, YOLO11-Nano achieved an mAP50 of 88.7% and an mAP50:95 of 56.0% with 2.59 million parameters and 6.4 GFLOPs, whereas YOLOv10-Nano achieved 85.9% and 56.5%, respectively, with 2.70 million parameters and 8.4 GFLOPs. Although YOLOv12-Nano produced slightly lower accuracy than YOLO11-Nano and YOLOv10-Nano, it required fewer parameters and lower computational complexity. This trade-off indicates that the lightweight architecture effectively preserves detection performance while improving computational efficiency, making it suitable for deployment on resource-constrained devices.

The competitive performance of YOLOv12-Nano can be attributed to its efficient feature extraction and multi-scale detection strategy, which enable the model to capture both global and local features while maintaining a compact architecture. However, its lightweight design also limits feature representation, particularly when the human body is partially occluded or occupies only a small region of the image. These conditions may reduce localization accuracy and increase the likelihood of missed detections. Figure 3 illustrates the detection results under different viewpoints, illumination conditions, and background environments. The model successfully detects most fall events, including cases with partial body visibility and challenging viewing angles.



Fig. 3. Fall detection results using the YOLOv12-Nano model.

This performance suggests that the learned feature representations generalize well across diverse environmental conditions. Nevertheless, detection performance is affected by factors such as severe occlusion, poor illumination, motion blur, and cluttered backgrounds, which reduce the visibility of body posture and make feature discrimination more difficult. Despite the overall satisfactory performance, detection errors remain. Some fall events generate multiple bounding boxes, while others are only partially detected or missed entirely. These errors are mainly caused by ambiguous body poses, incomplete object visibility, and the limited representation capability of the lightweight network. Improving feature representation and suppressing redundant predictions could further enhance detection accuracy in real-world applications.

IV. CONCLUSION

This study demonstrates the capability of the YOLOv12-Nano architecture to perform real-time fall detection under various environmental conditions. The model achieved a mean Average Precision (mAP50) of 86.6% and mAP50:95 of 55.5%, with 2.52 million parameters and a computational complexity of 6.0 GFLOPs, indicating its ability to maintain a balance between detection accuracy and computational efficiency. The experimental results also show that YOLOv12-Nano can handle multiple challenges in fall detection, including subjects that are only partially visible, challenging viewpoints, suboptimal lighting conditions, and variations in fall scenarios. However, some detection errors were still observed, such as bounding box overlaps or partial detections. Therefore, future research can focus on improving the attention mechanism or other image processing strategies to enhance precision. Overall, this study makes a significant contribution to the development of lightweight deep learning models for real-time human fall detection, while also providing a solid foundation for the implementation of safety and monitoring systems in real-world environments, particularly on devices with limited computational resources.

REFERENCES

- [1] United Nations, World Population Prospects 2022: Summary of Results, Department of Economic and Social Affairs, Population Division, 2022.
- [2] United Nations Department of Economic and Social Affairs, World Population Ageing 2020 Highlights: Living Arrangements of Older Persons, UNDESA, 2020.
- [3] World Health Organization, World Report on Ageing and Health, WHO Press, 2015.
- [4] World Health Organization, Falls: Key Facts, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/falls>.
- [5] A. Kaur, "State-of-the-art fall detection techniques with emphasis on vision-based approaches," *Sensors & Imaging*, 2025. (Review article covering methods and limitations)
- [6] Vision-based Human Fall Detection Systems: A Review, *Journal of Visual Communication and Image Representation*, 2024. (Ulasan menyeluruh tentang teknik berbasis visi dan dataset)
- [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. ICLR*, 2015.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [9] B.-H. Wang, J. Yu, K. Wang, X.-Y. Bao, and K.-M. Mao, "Fall detection based on dual-channel feature integration," *IEEE Access*, vol. 8, pp. 103443–103453, 2020, doi: 10.1109/ACCESS.2020.2999503.
- [10] A. Raza, M. H. Yousaf, and S. A. Velastin, "Human fall detection using YOLO: A real-time and AI-on-the-edge perspective," *2022 Int. Conf. on Pattern Recognition Systems (ICPRS)*, pp. 1–6, doi: 10.1109/ICPRS54038.2022.9854070.
- [11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proc. CVPR*, 2016.
- [12] C. Wang et al., "YOLOv12: Enhancing Real-Time Object Detection with Efficient Feature Aggregation," *arXiv preprint arXiv:2403.19241*, 2024.
- [13] Y. Tian, C. Yu, Y. Sun, L. Yang, W. Ren, X. Cao, Y. Qiao, and L. Zhang, "YOLOv12: Attention-Centric Real-Time Object Detectors," *arXiv preprint arXiv:2502.12524*, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2502.12524>.
- [14] Kan, X., Zhu, S., Zhang, Y., & Qian, C. (2023). A Lightweight Human Fall Detection Network. *Sensors*, 23(22), 9069. <https://doi.org/10.3390/s23229069>.
- [15] Y. Qin, W. Miao, and C. Qian, "A High-Precision Fall Detection Model Based on Dynamic Convolution in Complex Scenes," *Electronics*, vol. 13, no. 6, p. 1141, 2024. doi: 10.3390/electronics13061141.
- [16] H. Hwang, D. Kim, and H. Kim, "FD-YOLO: A YOLO Network Optimized for Fall Detection," *Applied Sciences*, vol. 15, no. 1, p. 453, 2025, doi: 10.3390/app15010453.
- [17] Wang, Y., Chi, Z., Liu, M., Li, G., & Ding, S. (2023). High-Performance Lightweight Fall Detection with an Improved YOLOv5s

- Algorithm. Machines, 11(8), 818. <https://doi.org/10.3390/machines11080818>.
- [18] Yang, Z., Tsui, B., Ning, J., & Wu, Z. (2024). Falling Detection of Toddlers Based on Improved YOLOv8 Models. *Sensors*, 24(19), 6451. <https://doi.org/10.3390/s24196451>.
- [19] J. Yang, Y. J. Jin and R. Y. C. Kim, "Multi-Person Fall Detection Using Data Assimilation Method With Kalman Filter," in *IEEE Access*, vol. 13, pp. 104692-104705, 2025, doi: 10.1109/ACCESS.2025.3578683.
- [20] D. Mohan et al., "Artificial Intelligence and IoT in Elderly Fall Prevention: A Review," *IEEE Sensors Journal*, vol. 24, no. 4, pp. 4181–4193, Feb. 2024, doi: 10.1109/JSEN.2023.3344605.