Fast and Lightweight Real-Time FallDetection System

[1]Chenzhe Shi and [2,*]Yue Kan,
[1,2]School of Mechanical and Power Engineering, Henan Polytechnic University, Jiaozuo, China
[*]Corresponding author

Abstract: Accurately identifying fall incidents in images or videos can significantly reduce the response time for assisting affected individuals. When manual monitoring cannot cover all video feeds in real time, fall detection algorithms can automatically screen for anomalous events. Existing methods are constrained by limitations such as homogeneous fall scenarios, insufficient fall-like activities, and low-resolution fall images. This paper introduces Multi-Variate Fall Detection Data (MVFDD), a novel and comprehensive fall detection dataset, along with a lightweight algorithm named FYOLO. An Identity Former block incorporating a Convolutional Gated Linear Unit (CGLU) has been introduced into the Cross Stage Partial Network of YOLOv10n, enabling channelwise feature modulation while reducing computational redundancy. The neck network leverages pinwheel-shaped convolution and frequency aware feature fusion, forming an Feature Pyramid Network (FPN) and Path Aggregation Network(PAN) structure to ensure effective detection of targets at various distances and sizes, thereby adapting to different camera perspectives and focal lengths. A Fall Distance Intersection over Union (FDIoU) loss is proposed for the first time, which enhances robustness to sample imbalance while ensuring semantic alignment with postural characteristics.

Keywords: *Deep learning, YOLOV10n, Lightweight, Fall detection*

I. Introduction

As the aging population increases, more individuals will be at risk of falling, and falls among older people are one of the most common and serious health concerns. By 2050, people older than 65 years are estimated to account for 16% of the population [1]. Falls pose a significant health threat due to their high incidence and mortality rates, substantial medical costs, and the serious complications that can follow.

The application value of computer vision and deep learning in detecting falling humans is becoming increasingly significant.In recent years, deep learning technology has achieved significant breakthroughs in the field of computer vision, particularly in object detection.

The Real-Time Detection Transformer (RT-DETR) model utilizes a high-efficiency hybrid encoder, featuring modules dedicated to intra-scale feature interaction and cross-scale context fusion [2]. Nano-Det achieves an excellent balance between model size, inference speed, and detection accuracy through its sophisticated lightweight design, making it an outstanding and practical choice for real-time object detection on mobile and embedded devices [3].

The You Only Look Once (YOLO) series is characterized by its continuous architectural evolution, consistently pushing the boundaries of real-time object detection by achieving an optimal balance between speed, accuracy, and computational efficiency. Stavros N. Moutsis et al. employed YOLOv8n, which was trained to recognize individuals in horizontal and vertical positions, along with an Support Vector Machine(SVM) classifier trained on a balanced dataset, to create an efficient lightweight system that runs smoothly on a Raspberry Pi 4 [4].

YOLOv10 achieves faster inference speed and higher detection accuracy by eliminating the need for Non-Maximum Suppression (NMS) and adopting a consistent dual label assignment strategy. This model significantly reduces computational overhead while maintaining high precision, making it particularly suitable for real-time application scenarios [5].

Despite these improvements in recognition accuracy, current network architectures continue to demonstrate insufficient parameter efficiency, high computational overhead, and limited inference speed, rendering them suboptimal for deployment on resource-constrained devices. Consequently, there is a critical need to develop a compact model architecture with low parameter counts and accelerated inference capabilities, tailored for applications in typical elderly home environments, while maintaining high detection accuracy.

To overcome the methodological and practical constraints of existing fall detection systems in general household settings, this research designed and implemented a fall detection system based on YOLOv10n.

II. MODELING AND SIMULATION DESCRIPTION

A. *The Backbone Network Architecture*

The Identity Former block was used[6]. It was integrated with the convolutional gated linear unit (CGLU) [7].This new block was then incorporated into the bottleneck of YOLOv10. A novel cross stage partial network cross stage partial fusion bottleneck with two convolutions bottleneck, an identity former block, and a convolutional Gated Linear Unit(C2f-IFCGLU) module, was developed.

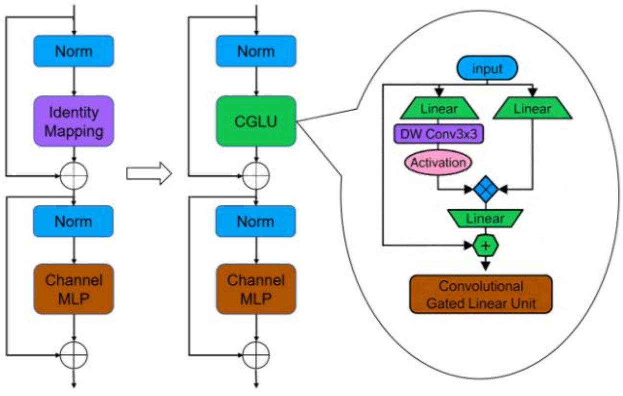

Fig.1 The structure of IFCGLU block

As illustrated in Fig.1, CGLU aims to bridge the gap between the GLU and the Squeeze-and-Excitation (SE) mechanism. In CGLU, a minimal 3×3 depth-wise separable convolution is incorporated into the gating branch of the GLU before the activation function. This depth-wise convolution operation enables each token to possess a unique gating signal based on its neighboring fine-grained image features. Replacing the traditional Channel Multilayer Perceptron (MLP) in the Identity Former block with CGLU replaces a spatially independent channel mixing approach with a hybrid method that integrates local spatial perception and a dynamic channel gating mechanism.

This enhancement strengthens the model's ability to capture local features and positional information, enabling more refined and dynamic channel-wise feature modulation while reducing computational redundancy to some extent in subsequent operations.

B. *The Neck Network Architecture*

The pinwheel-shaped convolution (PSConv) represents a dynamic, anisotropic, and scale-aware convolutional operation[8]. It decomposes a conventional K×K convolutional kernel intosub-kernels, each dedicated to extracting features along a specific directional orientation. For every pixel location within the input feature map, the network evaluates an attention weight vector comprising values based on the contextual features of the current pixel to determine the significance of each sub-kernel. The final convolutional output is obtained by aggregating the outputs of all sub-kernels through a weighted summation guided by these dynamically generated attention weights.

The core principle of Frequency-aware Feature Fusion (FreqFusion) is to optimize feature fusion through a frequency-aware mechanism. This approach first decomposes the feature maps into low-frequency components, which carry global semantics, and high-frequency components, which contain fine-grained textures, utilizing the Discrete Cosine Transform (DCT) [9]. Subsequently, a dual-path architecture is employed to process and integrate these high- and low-frequency features differentially. Finally, the processed components are reintegrated via spatially adaptive dynamic weighting maps, thereby preserving both detailed information and coherent semantic context within the fused features. The improved model is as shown in Fig. 2.

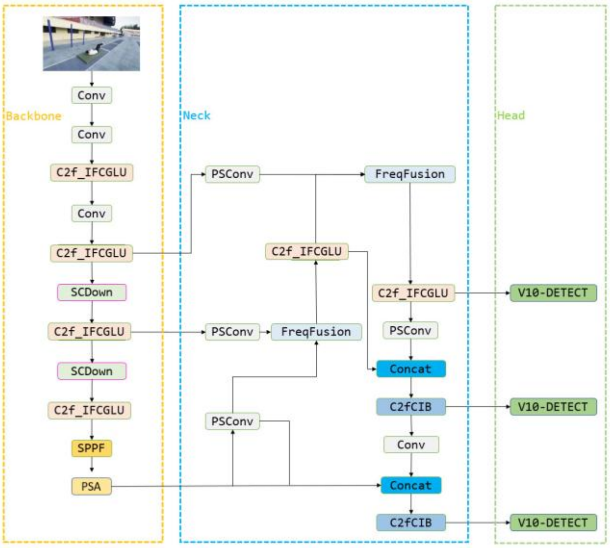

Fig.2 The structure of FYOLO

Through FreqFusion and multi-scale feature integration, the model can acutely capture the distinctive spatial-frequency characteristics and subtle postural changes associated with human falls, significantly reducing false positives such as misclassifying squatting as falling and minimizing missed detections. The frequency-domain processing, combined with the global modeling capability of Transformer, equips the model with strong resistance to real-world interference including lighting variations, partial occlusions, and complex backgrounds. The optimized Feature Pyramid Network (FPN)+Path Aggregation Network (PAN) [10] structure ensure effective detection of targets at various distances and sizes, adapting to different camera perspectives and focal lengths. PSConv reduces the channel count of high-level features, for instance from 1024 to 256, thereby decreasing the computational load and memory usage required for subsequent fusion.

C. *Design of F-DIoU Loss*

Conventional Intersection over Union (IoU) based loss functions, such as complete IoU (CIoU), generalized IoU (GIoU) and distance IoU (DIoU), exhibit significant limitations when applied to fall detection tasks. These shortcomings primarily stem from their insensitivity to bounding box aspect ratio and insufficient constraints on posture-specific features, both of which are critical for distinguishing between upright and fallen postures. Although CIoU introduces a shape alignment term to mitigate these issues, its reliance on an ambiguous aspect ratio formulation often leads to optimization instability. To address these deficiencies, we propose F-DIOU Loss, a novel loss function that preserves the stable convergence properties of DIoU while incorporating two key innovations: an adaptive focal weighting term using the square root of one minus IoU, and a pose-aware aspect ratio consistency constraint that explicitly penalizes posture mismatches between predicted and ground truth bounding boxes. This design ensures that the loss function is not only robust to sample imbalance but also semantically aligned with postural characteristics, thereby enhancing localization accuracy for fall detection applications.

The complete FDIOU Loss is the integration of the adaptively weighted DIoU term and the pose-aware constraintasin (1):

$$
\begin{aligned}
L_{\mathrm{fdiou}} = \sqrt{1-\mathrm{IoU}} \cdot \left(1-\mathrm{IoU}+\frac{\rho^2(\mathrm{bp,bgt})}{c^2}\right) \\
+ |O(\mathrm{bp})-O(\mathrm{bgt})| \\
\cdot \frac{\min(w_{\mathrm{p}},w_{\mathrm{gt}})+\min(h_{\mathrm{p}},h_{\mathrm{gt}})}{w_{\mathrm{gt}}+h_{\mathrm{gt}}} \qquad (1)
\end{aligned}
$$

where bp andbgt denote the predicted and ground-truth box centers. ρ is the Euclidean distance function, and c representsthe diagonal length of the smallest enclosing convexhull. When the posture labels of the predicted $O(\mathrm{bp})$and the ground-truth box $O(\mathrm{bgt})$ differ, the loss function applies a penalty.Posture indicator function $O(\mathrm{b})$ as shown in (2):

$$
O(\mathrm{b}) = \begin{cases} 1 & \text{if } w > h \\ 0 & \text{if } w \le h \end{cases} \#(2)
$$

This penalty is scaled based onthe normalized similarity between the predicted box dimensions w_{p}, h_{p} and ground-truth dimensionsw_{gt},h_{gt}.

The property of the square root function automatically assigns higher weights to hard samples and lower weights to easy ones, thereby fulfilling the original intention of focal loss while eliminating the need for tedious hyperparameter optimization. This adaptive gradient regulation mechanism enables the model to learn rapidly from all samples in the initial training phase and subsequently concentrate on challenging hard to regress samples in later stages. Moreover, it avoids assigning excessive weight to outlier samples, thereby enhancing training stability and ultimately leading to improved overall performance and more stable convergence. The posture-aware aspect ratio consistency constraint further enhances the loss function by explicitly encoding the critical biomechanical feature of human poses, where standing postures typically exhibit height-dominant bounding boxes while fallen postures exhibit width-dominant ones. Ultimately, these innovations lead to improved overall performance with more stable convergence.

III. Experimental section

To further enhance the complexity of fall samples, by designing human fall and fall-like activities, we collected data and created a corresponding dataset named Multi-Variate Fall Detection Data(MVFD). The dataset comprises 14 male and 7 female participants with significant variations in body type, clothing, and age. It covers six indoor and four outdoor falling scenarios, with four different falling directions: forward, backward, left, and right. The types of falls include single-person falls, simultaneous two-person falls, and falls occurring within crowds. Fall-like behaviors such as crawling, squatting, bending, lying down, side-lying, jumping, and push-ups are also included. Lighting conditions vary across three levels: adequate, low, and insufficient. Some examples of this data are shown in Fig.3.

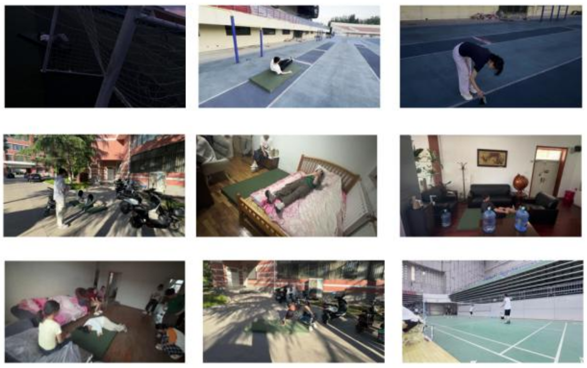

Fig.3 The details of MVFDD

The collected data was organized into a YOLO dataset, consisting of 11473 images in total.The dataset was stratified into training, validation, and test sets at a ratio of 8:1:1.

A. Set- up

Experiments were conducted on a system equipped with an Intel(R) Core(TM) i9-14900K processor and an NVIDIA GeForce Ray Tracing Texel eXtreme 4090 graphics card. The software stack utilized PyTorch 2.1.0 accelerated by CUDA 11.8, running on Python 3.9. To optimize model performance and prevent overfitting, a 150-epoch training schedule was employed.

B. Experimental result

Fig.4 shows the performance trajectories of Precision, Recall, and mean Average Precision (mAP) for the FYOLOmodel across 150 training epochs. All metrics improve progressively with each iteration until they plateau at an optimal level. The model's bounding box loss (box_loss), classification loss (cls_loss), and distribution focal loss (dfl_loss) exhibit a rapid and steady downward trend on both the training and validation sets. After approximately 50 epochs, these loss values tend to stabilize, and the validation loss curves closely align with the training curves, showing no obvious signs of overfitting. This demonstrates that the proposed lightweight model possesses exceptional learning capability and outstanding generalization performance.

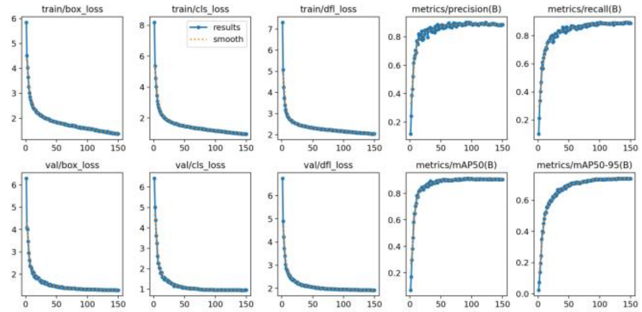

Fig.4 Performance trajectories of FYOLO

As shown in Fig.5, precision-recall curve of the FYOLO model approaches the top-right corner (1, 1), indicating that the model achieves both high precision and high recall, effectively identifying positive-class samples with low misjudgment rates. For the critical 'fall' category, the AP reaches 0.899, while for the 'non-fall' category, it peaks at 0.933. Additionally, the mAP50-95 metric shown in Figure 6 remains stable at a high level (around 0.7 or above). This demonstrates that the model is highly accurate not only in classifying target classes, but also in localizing bounding boxes.

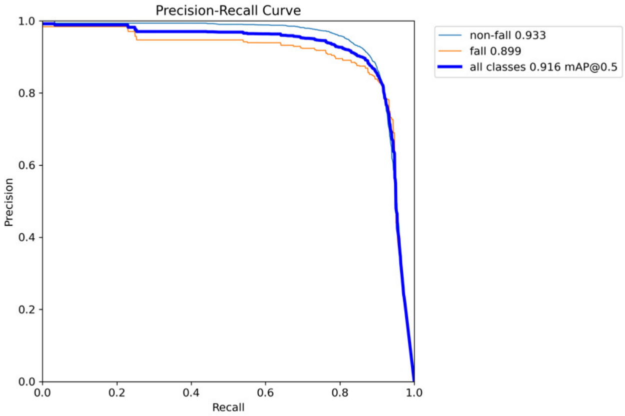

Fig.5 The P-R Curve of FYOLO

As shown in Fig.6, F1–confidence curve reflects the stability and effectiveness of model predictions by showing the variation in F1-scores under different confidence thresholds. The curve reaches the optimal F1-score of 0.88 at a confidence threshold of 0.427, and FYOLO demonstrates strong performance under high confidence levels.

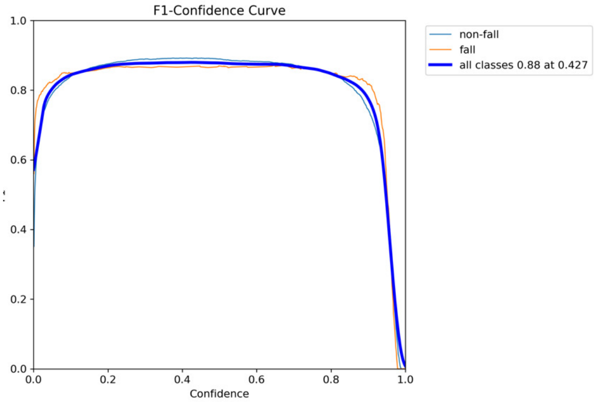

Fig.6 The F1 Curve of FYOLO

Additionally, the YOLOv10n baseline achieved 90.1% mAP50 and 73.1% mAP95 with 2.37M parameters and 6.7GFLOPs.The FYOLOmodel achieved the efficiency with only 1.68M parameters and 4.8GFLOPs,while attaining 91.6% mAP50, 75.4% mAP95.

CONCLUSION

This paper proposes Fall detection YOLO (FYOLO), a lightweight fall detection model. Its architecture builds upon YOLOv10n with the following innovations:

- Introduction of C2F-IFCGLU, which integrates the IdentityFormer Block and CGLU to form a novel block replacing the original bottleneck in the C2F module.
- Proposal of a new structure based on PSConv and Freqfusion to replace the original neck of YOLOv10n, resulting in reduced parameters and computational costs.
- Design of an F-DIoU loss function specifically for fall detection, further enhancing model accuracy.
- Creation of a new fall detection dataset encompassing diverse fall types, scenarios, and fall-like actions, which contributes to improved algorithm accuracy and generalization.

The novel lightweight YOLO model proposed in this paper significantly reduces network complexity while maintaining exceptionally high detection accuracy and stability. Achieving an mAP@0.5 of 91.6% and an F1 score of 0.88, it can accurately and efficiently perform the fall detection task. Consequently, it fully meets the dual requirements of real-time processing and high accuracy in real-world scenarios, demonstrating significant value for practical engineering applications.

References

[1] WHO. WHO Global Report on Falls: Prevention in Older Age[M]. World Health Organization, 2007.

[2] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.

[3] RangiLyu: Nanodet. https://github.com/RangiLyu/nanodet

[4] Moutsis S N, Tsintotas K A, Kansizoglou I, et al. Fall detection paradigm for embedded devices based on YOLOv8[C]//2023 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE, 2023: 1-6.

[5] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[6] Author, Book title, page numbers. Publisher, place Yu W, Si C, Zhou P, et al. Metaformer baselines for vision[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(2): 896-912.

[7] Shi D. Transnext: Robust foveal visual perception for vision transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 17773-17783.

[8] Yang J, Liu S, Wu J, et al. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(9): 9202-9210.

[9] Ahmed N, Natarajan T, Rao K R. Discrete cosine transform[J]. IEEE transactions on Computers, 2006, 100(1): 90-93.

[10] Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.

[11] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.