

Article

Research on Passenger Posture Prediction in Subway Station Based on Lightweight Improved YOLOv8-Pose

Ye Zhang ^{1,*}, Weidong Yan ^{1,*}, Tuo Shan ², Guoqi Liu ³ and Yuxuan Li ³

¹ School of Civil Engineering, Shenyang Jianzhu University, Shenyang 110168, China

² School of Electrical and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China

³ School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang 110168, China

* Correspondence: zy9067@sjzu.edu.cn (Y.Z.); ywd@sjzu.edu.cn (W.Y.); Tel.: +86-15940173298 (Y.Z.); +86-13940076600 (W.Y.)

Abstract

To address challenges including severe occlusion, large-scale variation, inaccurate key-point localization, and strict real-time constraints in subway surveillance scenarios, this paper proposes a lightweight human pose estimation framework based on an enhanced YOLOv8-Pose architecture. The proposed method introduces four key components: a DWR (Dilation-wise Residual) module to expand multi-scale receptive fields for improved joint context modeling, an HFERB (High-Frequency Enhancement Residual Block) to recover fine-grained structural details lost during downsampling, an OrthoNets-based orthogonal attention mechanism to reduce channel redundancy while preserving discriminative representations, and an MFEB (Multi-Scale Feature Enhancement Block) to enable adaptive feature fusion under varying crowd densities. Extensive experiments on the COCO 2017 dataset demonstrate that the proposed method achieves 87.9% mAP@0.5 and 64.9% mAP@0.5–0.95, improving the baseline YOLOv8-Pose by 1.7% and 1.0%, respectively, while reducing validation pose loss by 5.77%. After lightweight optimization with ONNX export and FP16 inference, the model is compressed to 3.99M parameters and achieves 76 FPS inference speed on the same hardware, significantly improving efficiency compared with the baseline, while achieving 38 FPS on the full model. Both configurations satisfy real-time constraints for subway surveillance deployment. Qualitative results on real subway surveillance videos further demonstrate strong robustness under occlusion and scale variation. Overall, the proposed method achieves a favorable trade-off between accuracy and efficiency, making it suitable for real-time human pose estimation in intelligent transportation systems.

Keywords: subway surveillance; human pose estimation; YOLOv8-Pose; DWR; HFERB; OrthoNets; MFEB; lightweight network



Academic Editor: John A. Kalomiros

Received: 25 June 2026

Revised: 6 July 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As a fundamental component of modern urban transportation systems, urban rail transit plays a critical role in ensuring operational safety and management efficiency. In densely populated environments such as subway station concourses and train carriages, real-time human pose estimation based on surveillance cameras provides essential technical support for applications including abnormal behavior detection (e.g., falling, fighting, and crowd aggregation), passenger flow analysis, security assistance, and human–computer interaction systems such as contactless ticketing [1,2].

However, subway surveillance scenarios are characterized by severe occlusions, large variations in human scale, complex illumination conditions, and strict real-time constraints, which collectively pose significant challenges to both accuracy and efficiency of pose estimation models [3,4].

Human pose estimation aims to localize human keypoints and infer underlying skeletal structures, and has become a core technology in intelligent surveillance and behavior understanding systems [5]. Recent research trends focus on improving accuracy, enhancing real-time constraints, and reducing model complexity to facilitate deployment on edge devices [6]. In practical subway environments, robust models must simultaneously handle dense occlusions and multi-scale pedestrian distributions while meeting real-time constraints [7–9].

Existing methods can be broadly categorized into bottom-up and top-down paradigms. Bottom-up approaches such as OpenPose [10] utilize Part Affinity Fields to associate keypoints across multiple individuals. Although efficient in multi-person scenarios, they often suffer from information loss due to repeated downsampling, leading to degraded performance in crowded scenes. Top-down methods such as HRNet [11] maintain high-resolution representations throughout the network and achieve superior accuracy; however, they incur high computational costs and are sensitive to occlusion and background interference. Simple Baseline [12] simplifies the pipeline but lacks robust multi-scale modeling capability, resulting in reduced accuracy under small-object conditions.

To further address occlusion challenges, AlphaPose [13] introduces a multi-stage refinement strategy, but its complex pipeline limits real-time performance and hinders real-time deployment. YOLO-based approaches, such as YOLO-Pose [14] and YOLOv8-Pose [15], integrate detection and pose estimation into a unified framework, achieving a favorable trade-off between accuracy and efficiency. However, their ability to model fine-grained multi-scale relationships remains limited in dense crowd scenarios.

Transformer-based approaches such as TokenPose [16] improve global modeling capability via self-attention mechanisms but suffer from high computational complexity. Channel attention mechanisms, including OrthoNets [17], reduce redundancy while preserving discriminative features. Additionally, multi-scale feature enhancement strategies [18–20] demonstrate strong effectiveness in improving robustness under occlusion and scale variation.

Despite these advances, existing methods still face a fundamental trade-off among accuracy, robustness, and efficiency. YOLOv8-Pose achieves a relatively balanced performance but remains limited in handling severe occlusion and small-scale joint localization in dense environments.

To address these challenges, this paper proposes a lightweight and efficient pose estimation framework based on an improved YOLOv8-Pose architecture. The proposed method enhances feature representation from four complementary perspectives: multi-scale contextual modeling, high-frequency detail preservation, channel redundancy reduction, and adaptive feature fusion.

The main contributions of this work are summarized as follows:

1. A lightweight YOLOv8-Pose-based pose estimation framework is proposed for dense subway surveillance scenarios, achieving a balance between accuracy and efficiency.
2. A DWR module is introduced to enhance multi-scale contextual modeling via dilated residual convolutions.
3. An HFERB module is designed to recover high-frequency structural details and improve small-joint localization accuracy.
4. An orthogonal attention mechanism based on OrthoNets is adopted to reduce channel redundancy while preserving discriminative representations.

5. An MFEB module is proposed to enable adaptive multi-scale feature fusion under varying crowd densities.

2. Experiments and Methods to Improve YOLOv8-Pose Model

YOLOv8-Pose is a model within the YOLOv8 framework that simultaneously performs object bounding box detection and human keypoint estimation. It consists of four main components, namely the input module, backbone network, neck network, and detection head, as illustrated in Figure 1.

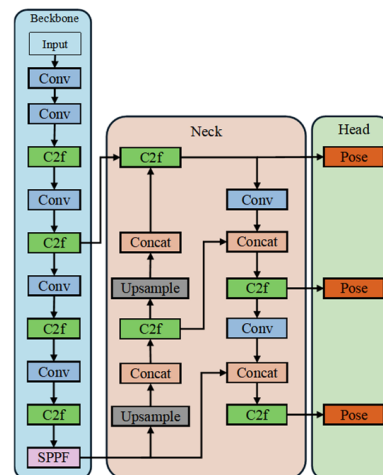


Figure 1. Network structure of YOLOv8-Pose.

To address the insufficient accuracy of existing YOLOv8-Pose models in complex scenarios involving small objects, occlusions, and multi-scale human pose estimation, while meeting real-time constraints, this paper innovatively introduces four core modules into the backbone and neck of YOLOv8-Pose: the DWR dilated residual module, OrthoNets orthogonal attention, HFERB, and MFEB. These modules are designed to address key challenges including receptive field expansion, channel redundancy reduction, high-frequency detail restoration, and adaptive multi-scale feature fusion, respectively. Together, they form a lightweight yet high-accuracy pose estimation framework. The overall network architecture is illustrated in Figure 2.

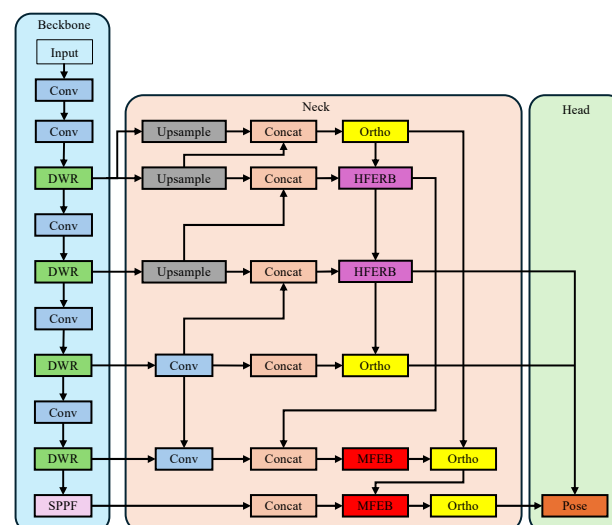


Figure 2. Structure diagram of the improved YOLOv8-Pose model.

2.1. DWR Module

In human pose estimation tasks, accurately modeling spatial dependencies among different joints is crucial. For example, the distance between fingers and wrists is much smaller than the spatial span of the torso, which requires the network to possess the ability to perceive multi-scale spatial relationships. However, conventional convolutional layers are limited by fixed receptive fields and struggle to effectively capture such multi-scale contextual information.

To this end, a DWR module is introduced into the shallow feature extraction stage of the backbone network (i.e., the P2 stage), as illustrated in Figure 3.

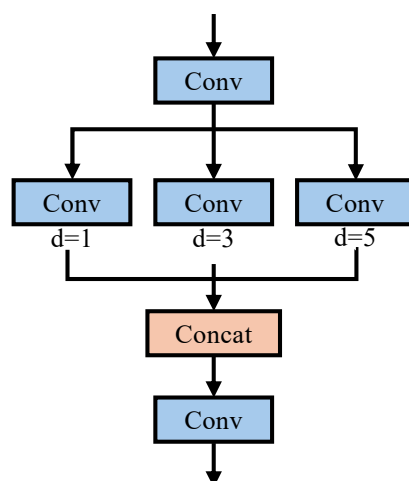


Figure 3. Network structure of DWR dilation residual module.

The core idea of the proposed module is to capture multi-scale contextual information by applying parallel 3×3 convolutions with different dilation rates, thereby enlarging the receptive field without significantly increasing the number of parameters. Given an input feature map $X \in R^{H \times W \times C}$, a standard 3×3 convolution is first applied to reduce the channel dimension to half, producing a compressed feature representation $X_1 \in R^{H \times W \times C/2}$:

$$X_1 = \text{Conv}_{3 \times 3}(X) \quad (1)$$

Subsequently, X_1 is fed into three parallel 3×3 dilated convolutions with dilation rates $d \in \{1, 3, 5\}$, generating multi-scale feature representations with different receptive fields:

$$F_d = \text{Conv}_{3 \times 3, d}(X_1), d \in \{1, 3, 5\} \quad (2)$$

The corresponding receptive fields are approximately 3×3 , 9×9 , and 15×15 , respectively, which align well with the varying spatial distances between human joints in pose estimation. The three feature maps are then concatenated along the channel dimension, followed by a 1×1 convolution for feature fusion. Finally, a residual connection is applied to obtain the output of the module:

$$Y = \text{Conv}_{1 \times 1}(\text{Concat}(F_1, F_3, F_5)) + X \quad (3)$$

2.2. HFERB Module

In the neck and head upsampling pathways of pose estimation networks, repeated downsampling operations often lead to the loss of high-frequency details such as joint edges and skeletal contours. This issue is particularly detrimental to the localization of small-scale joints, including wrists and ankles. To address this problem, inspired by the explicit high-frequency enhancement strategy proposed in CRAFT (ICCV 2023) [21], we

design the High-Frequency Enhancement Residual Block (HFERB), which is inserted after the P3 upsampling fusion stage in the neck. The network architecture is illustrated in Figure 4.

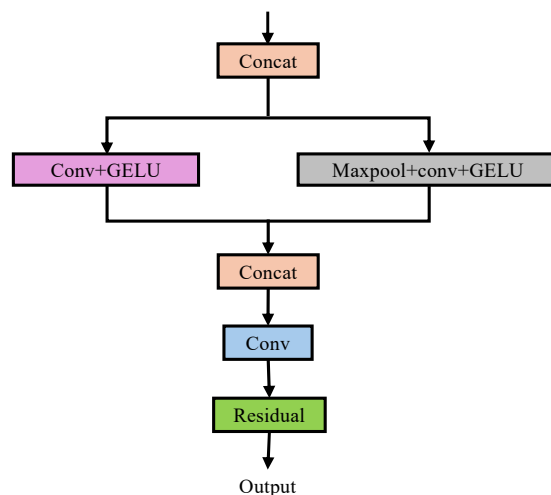


Figure 4. Network structure of High-Frequency Enhancement Residual Block.

First, the input feature map $X \in \mathbb{R}^{H \times W \times C}$ is split evenly along the channel dimension into two parts, X_l and X_h , corresponding to the low-frequency and high-frequency branches, respectively.

In the low-frequency branch, a standard 3×3 convolution is applied to extract smooth local semantic features. In the high-frequency branch, a max-pooling operation is first performed. Since max pooling inherently emphasizes local intensity variations, it can effectively highlight edge responses and enhance boundary information of joints. Then, a 1×1 convolution is used for channel transformation.

The outputs of the two branches are concatenated along the channel dimension and further fused by a 1×1 convolution. Finally, a residual connection is introduced to combine the fused features with the original input, preserving original information while injecting enhanced high-frequency components. The overall process can be formulated as:

$$Y = \text{Conv}_{1 \times 1}(\text{Concat}(\text{Conv}_{3 \times 3}(X_l), \text{Conv}_{1 \times 1}(\text{MaxPool}(X_h)))) + X \quad (4)$$

This module is lightweight, introducing only approximately 3% additional parameters, while significantly improving the localization accuracy of small-scale joints (e.g., wrists, ankles, and eyes). The improvement is particularly pronounced under low-resolution inputs and in crowded scenarios with severe pose occlusions.

2.3. Orthogonal Attention Module (OrthoNets-Based)

Channel attention mechanisms (e.g., SE blocks) typically compress spatial information via global average pooling. However, this compression process often introduces information redundancy across channels, which may cause discriminative joint features to be suppressed. To address this issue, this paper introduces an improved orthogonal attention module based on the OrthoNets framework (CVPR 2023 [20]), referred to as Gram–Schmidt Orthogonal Attention. The module is deployed at the P4 stage of the neck network to enhance channel selection capability during feature fusion, as illustrated in Figure 5.

The module first constructs a set of orthogonal constant filters W_o , which satisfy the orthogonality constraint $W_o^T W_o = I$, where I denotes the identity matrix. The input

feature map $X \in R^{H \times W \times C}$ is then iteratively projected and compressed into a 1×1 spatial representation $Z \in R^{1 \times 1 \times C}$. This process can be formulated as:

$$Z = W_o \times X \quad (5)$$

Due to the orthogonality constraint, the projection reduces channel redundancy while preserving discriminative information across channels, preventing cross-channel interference during compression.

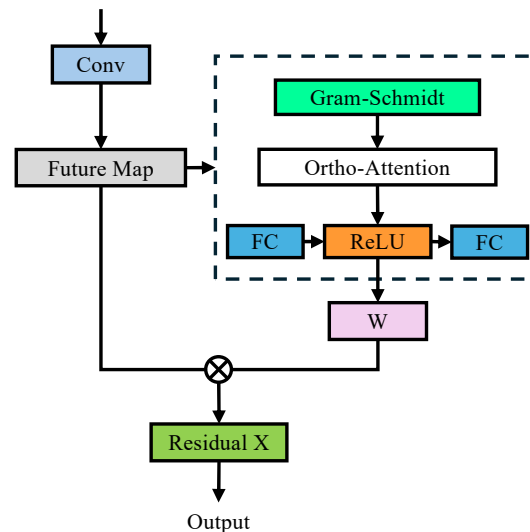


Figure 5. OrthoNets orthogonal attention mechanism.

Subsequently, similar to the SE mechanism, the compressed features are passed through a two-layer fully connected network to generate channel-wise attention weights $W \in R^C$:

$$W = \sigma(FC_2(ReLU(FC_1(Z)))) \quad (6)$$

Finally, the learned attention weights are applied to the original feature map via channel-wise multiplication to obtain the refined output feature:

$$Y = W \otimes X \quad (7)$$

where $\otimes$ denotes element-wise channel-wise multiplication.

Compared with conventional SE modules, the proposed orthogonal compression preserves more discriminative structural information related to joint geometry and relative spatial relationships. Experimental results demonstrate that this module is particularly effective in improving keypoint detection performance under occlusion-heavy scenarios. In practice, orthogonal constraints are enforced through Gram–Schmidt-based initialization followed by learnable fine-tuning.

2.4. MFEB Module

To address the challenges in crowded occlusion scenarios, where human body scales vary significantly and a single receptive field is insufficient to model complex limb structures, this paper draws inspiration from multi-scale feature enhancement and adaptive fusion strategies in medical image segmentation. By improving the conventional ASPP structure, we propose the Multi-Scale Feature Enhancement Block (MFEB) [22], an adaptive multi-scale enhancement module. The module integrates a cascaded dilated residual structure with a SE-based adaptive weighting mechanism, enabling the network to dynamically

select the most informative scale features according to the density characteristics of the input scene. The overall architecture is illustrated in Figure 6.

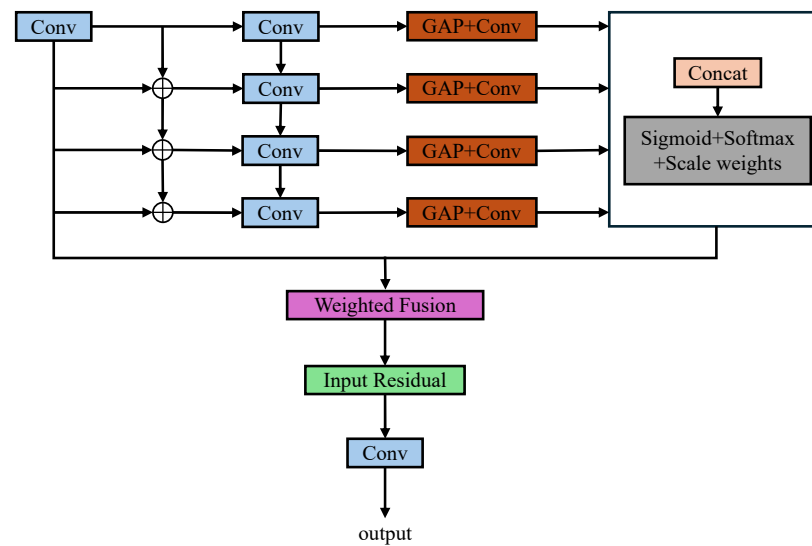


Figure 6. Network structure of Multi-Scale Feature Enhancement Block.

The module employs parallel atrous convolutions with dilation rates of 1, 2, 4, and 8 to progressively capture multi-scale limb representations. Each branch incorporates a residual connection with the original input to prevent feature degradation. The formulation is given as:

$$X_i = \text{Conv}_{d=i}(X_{in}) + X_{in}, \quad i \in \{1, 2, 4, 8\} \quad (8)$$

To enable adaptive multi-scale fusion, a channel descriptor extraction branch is first employed to generate feature representations, followed by a Softmax normalization to produce competition-aware attention weights across different feature streams.

$$w_i = \text{Softmax}(\text{SE}(X_i)) \quad (9)$$

Finally, the multi-scale features are aggregated through weighted summation to achieve global adaptive feature enhancement:

$$X_{out} = \sum w_i \cdot X_i \quad (10)$$

The MFEB module is deployed at the P4 deep feature stage. It adaptively aligns representations across scales ranging from small joints to large body structures, effectively improving feature stability and detection recall in occluded and densely populated scenarios.

2.5. Lightweight Deployment Optimization

Considering that pose estimation algorithms are often deployed on edge devices such as robots, smartphones, and surveillance cameras in practical applications, this study further performs lightweight deployment optimization on the improved model. Specifically, the trained PyTorch (2.0.1) weights are converted into the ONNX (Open Neural Network Exchange) format [23] to ensure cross-platform compatibility. In addition, half-precision inference is adopted, converting both model parameters and intermediate activations from 32-bit floating-point (FP32) to 16-bit floating-point (FP16) [24].

This strategy reduces memory consumption and improves inference efficiency under compatible hardware and software environments, while maintaining comparable pose estimation accuracy. Furthermore, ONNX-based models provide compatibility with

commonly used inference engines such as TensorRT [25] and OpenVINO [26], enabling potential acceleration on different hardware platforms. However, this study focuses on algorithmic performance evaluation, and hardware-specific deployment benchmarks are left for future work.

3. Experimental Results and Analysis

3.1. Experimental Environment and Dataset

The experimental environment is built on the PyTorch deep learning framework under a Windows 10 (Redmond, WA, USA) operating system, implemented using Python 3.9 (Beaverton, OR, USA). The hardware configuration includes an Intel Core i7-10700 CPU (Santa Clara, CA, USA) and an NVIDIA RTX 3060 GPU (Santa Clara, CA, USA) with 16 GB of memory. The CUDA version is 11.0. The model is trained for 100 epochs with a batch size of 32. The initial learning rate is set to 0.01, and the input image resolution is 320×320 .

The COCO 2017 dataset is adopted in this study, which is widely used for human pose estimation tasks. The training set contains 118,287 images, while the validation set includes 5000 images. In practical pose estimation evaluation, images without valid human instances or complete keypoint annotations are excluded. After this pose-validity filtering process, 2346 images are retained as the effective evaluation subset. This subset is consistently used across all experiments to ensure fair comparison under identical evaluation settings. The dataset covers diverse scenarios including single-person, multi-person, large-scale and small-scale human instances, enabling comprehensive evaluation of model robustness.

In addition, to further evaluate the model performance in real-world surveillance scenarios, a video dataset collected from Line 1 of the Shenyang Metro in China is used as an additional test set. This dataset consists of 105 surveillance videos recorded in station platforms, concourses, and transfer corridors. All videos have a resolution of 704×576 , are stored in MP4 format without audio, and recorded at 25 frames per second in 24-bit RGB format. Unlike the COCO-based evaluation, this video dataset is used only for qualitative and supplementary quantitative validation under real subway surveillance conditions, and is not involved in model training or COCO-based benchmarking. Representative frames are extracted for evaluation to demonstrate the generalization ability of the proposed method in complex crowded environments.

3.2. Evaluation Metrics

In this study, mAP@0.5, mAP@0.5–0.95, Precision, Recall, and validation pose loss (val pose_loss) are adopted as evaluation metrics to comprehensively assess the performance of the proposed method. The corresponding formulations are given as follows:

$$mAP50 = \frac{1}{K} \sum_{k=1}^K AP_k(OKS \geq 0.5) \quad (11)$$

$$mAP50-95 = \frac{1}{10} \sum_{t=0.5}^{0.95} mAP(t) \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$pose_loss = \frac{1}{N \cdot K} \sum_{n=1}^N \sum_{k=1}^K v_{n,k} \left[(x_{n,k} - \hat{x}_{n,k})^2 + (y_{n,k} - \hat{y}_{n,k})^2 \right] \quad (15)$$

where K denotes the total number of human keypoints, and AP_k represents the average precision of the k -th keypoint. t denotes the Object Keypoint Similarity (OKS) threshold,

and $mAP(t)$ represents the mean Average Precision computed under threshold t . TP (True Positives) denotes correctly detected pose instances, FP (False Positives) denotes incorrectly detected instances, and FN (False Negatives) denotes missed detections. N denotes the total number of samples in the validation set. $x_{n,k}$, $y_{n,k}$ and $\hat{x}_{n,k}$, $\hat{y}_{n,k}$ represent the ground-truth and predicted coordinates of the k -th keypoint in the n -th sample, respectively. $v_{n,k}$ is the visibility weight of the keypoint, taking a value of 1 if the keypoint is visible and 0 otherwise.

3.3. Analysis of Experimental Results

3.3.1. Ablation Study

To evaluate the contribution of each proposed module to the overall model performance, a series of ablation experiments are conducted. The corresponding results are summarized in Table 1.

Table 1. Results of ablation experiments.

DWR	OrthoNets	HFERB	MFEB	Full Configuration	mAP@0.5	mAP@0.5–0.95	Precision	Recall	Params (M)
×	×	×	×	×	0.862	0.639	0.865	0.762	11.6
✓	×	×	×	×	0.871	0.642	0.867	0.766	12.4
✓	✓	×	×	×	0.874	0.644	0.869	0.769	13.0
✓	✓	✓	×	×	0.876	0.646	0.871	0.771	13.4
✓	✓	✓	✓	×	0.877	0.647	0.872	0.773	14.2
✓	✓	✓	✓	✓	0.879	0.649	0.874	0.775	14.8

Table 1 summarizes the contribution of each module. Results show that each component consistently improves model performance under minimal computational overhead. In particular, DWR effectively enhances contextual modeling, OrthoNets reduces channel redundancy, HFERB improves fine-grained localization, and MFEB contributes the most significant gain through adaptive multi-scale fusion.

Overall, the full configuration achieves a unified improvement over the baseline, with consistent gains in accuracy, precision, and recall, while maintaining manageable model complexity. To further illustrate the effectiveness of the proposed method, three representative detection results are visualized in Figure 7. As shown, significant performance differences can be observed between the proposed method and the baseline YOLOv8-Pose in crowded scenes, occlusion scenarios, and small-scale human instances. The baseline model tends to suffer from missing keypoints, inaccurate joint localization, and limb ambiguity in densely overlapping regions. For distant or small-scale persons, it often produces incomplete joint representations and structurally inconsistent skeletons, failing to accurately capture human pose configurations.

In contrast, the proposed method, benefiting from multi-scale feature enhancement, high-frequency detail restoration, and channel redundancy suppression, can effectively distinguish overlapping human instances and accurately localize fine-grained joints such as wrists and ankles. Moreover, under heavy occlusion, the improved model is still capable of producing complete and anatomically consistent skeletal structures, demonstrating superior robustness and feature adaptability compared to the baseline model. The visualization results further validate that the proposed modules significantly enhance pose estimation accuracy and robustness in complex and dense scenarios.

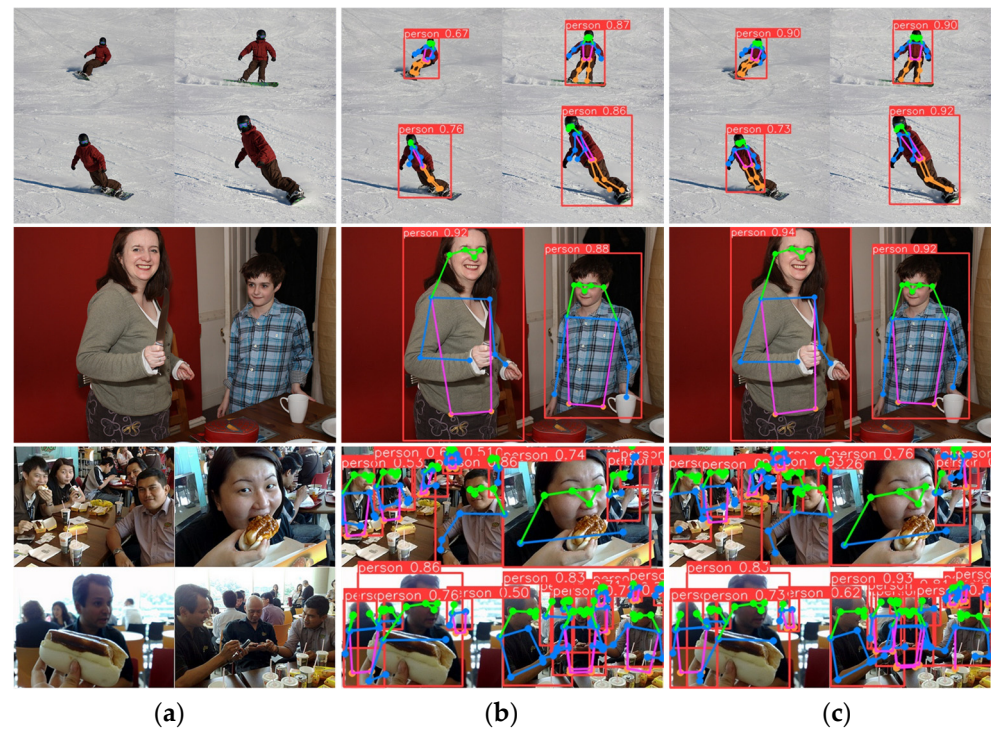


Figure 7. Comparison before and after improvement: (a) label; (b) YOLOv8-Pose; (c) Ours.

3.3.2. Comparative Experiments

To further validate the effectiveness of the proposed method, comparative experiments are conducted against several state-of-the-art human pose estimation approaches. For a fair comparison, all methods are evaluated on the COCO 2017 dataset under their standard settings. To reduce the influence of randomness, all reported results are obtained by averaging multiple independent runs under the same experimental configuration. The results are summarized in Table 2.

Table 2. Performance comparison with representative pose estimation methods.

Model	Backbone	InputSize	mAP@0.5	FLOPs (G)	Params (M)
Higher HRNet [11]	HRNet-W32	512×512	0.870	46.58	28.5
OpenPose [10]	VGG19	224×224	0.849	160.0	-
DeepPose [27]	ResNet-101	256×192	0.824	7.7	23.6
Efficient HRNet [28]	HRNet	480×480	0.826	14.2	16.0
YOLOv5s-Pose [14]	DarkNet-53	640×640	0.822	16.5	12.6
PCE_LitePose [29]	DarkNet-53	640×640	0.857	25.0	8.47
RTMPose-x [30]	CSPNeXt	384×288	0.788	17.22	49.43
YOLOv8-Pose [15]	Conv-C2f-SPPF	640×640	0.862	30.2	11.6
Ours	Improved Conv-C2f-SPPF	640×640	0.883	~38.5	14.8
Ours	Improved Conv-C2f-SPPF	320×320	0.879	~9.6	14.8
Ours (lightweight)	Lightweight Improved Conv-C2f-SPPF	640×640	0.873	~10.4	3.99
Ours (lightweight)	Lightweight Improved Conv-C2f-SPPF	320×320	0.870	~2.6	3.99

As shown in Table 2, the proposed method achieves an mAP@0.5 of 87.9% at an input resolution of 320×320 , outperforming all compared methods with lower computational cost. Compared with YOLOv8-Pose, PCE_LitePose, and HigherHRNet, the proposed approach consistently improves accuracy while maintaining a more efficient input configuration. To ensure a fair and consistent comparison, all baseline results are reported under

their officially published settings on the COCO2017 dataset. Notably, even when compared with HigherHRNet using a higher input resolution of 512×512 and a larger number of parameters, the proposed method still achieves superior performance, demonstrating the effectiveness of the proposed feature enhancement modules. Furthermore, the lightweight variant reduces the model size to 3.99M parameters while maintaining competitive accuracy, with only a marginal performance degradation. Overall, the results demonstrate that the proposed method achieves a favorable trade-off between accuracy and efficiency, making it suitable for resource-constrained deployment scenarios.

3.3.3. Loss Variation and Convergence Analysis

To comprehensively evaluate the effectiveness of the proposed improvements, the loss values and convergence behavior of the proposed model are compared with the baseline YOLOv8-Pose on the COCO 2017 dataset. The loss metrics directly reflect the discrepancy between model predictions and ground-truth annotations, serving as a fundamental indicator of fitting quality, generalization ability, and training stability. In addition, the number of training epochs required to reach different accuracy thresholds reflects the learning speed, feature extraction capability, and performance ceiling of the model. The comparative results are summarized in Table 3.

Table 3. Comparison of loss values and convergence properties.

Indicator Types	YOLOv8-Pose	Ours	Improvement
Validation Pose Loss	3.2969	3.1066	−5.77% (decrease)
Validation Box Loss	0.9656	0.9383	−2.83% (decrease)
Epochs to reach Pose mAP@0.5 ≥ 0.60	Epoch 20	Epoch 31	Baseline faster by 11 epochs
Epochs to reach Pose mAP@0.5–0.95 ≥ 0.35	Epoch 77	Epoch 70	Ours faster by 7 epochs
Epochs to reach Pose mAP@0.5 ≥ 0.70	Epoch 89	Epoch 79	Ours faster by 10 epochs

As shown in Table 3, the proposed method demonstrates lower validation pose loss and bounding box loss compared with the baseline, indicating improved localization accuracy and better generalization capability.

Although the baseline converges faster in early epochs, the proposed model exhibits stronger mid-to-late stage convergence behavior, achieving higher performance thresholds earlier and avoiding saturation effects observed in the baseline.

3.4. Visualization Results

To further validate the qualitative improvements beyond the quantitative metrics reported in Section 3.3, a visual comparison between the baseline YOLOv8-Pose and the proposed method is conducted on subway surveillance video frames. Figures 8 and 9 illustrate pose estimation results under two typical scenarios, namely sparse passenger flow and dense passenger flow, in station hall and platform environments.

In sparse crowd scenarios, the baseline model frequently suffers from missing joints and skeletal misalignment, particularly for distant and small-scale pedestrians. This is mainly attributed to repeated downsampling operations, which lead to the loss of high-frequency details, as well as a limited receptive field that fails to capture fine-grained joint structures at long distances. In contrast, the proposed method benefits from the HFERB module, which restores high-frequency boundary information, and the DWR module, which provides multi-scale receptive fields. As a result, it significantly improves skeleton completeness and detection recall for small-scale targets, producing more continuous and anatomically consistent pose estimations closely aligned with ground-truth distributions.

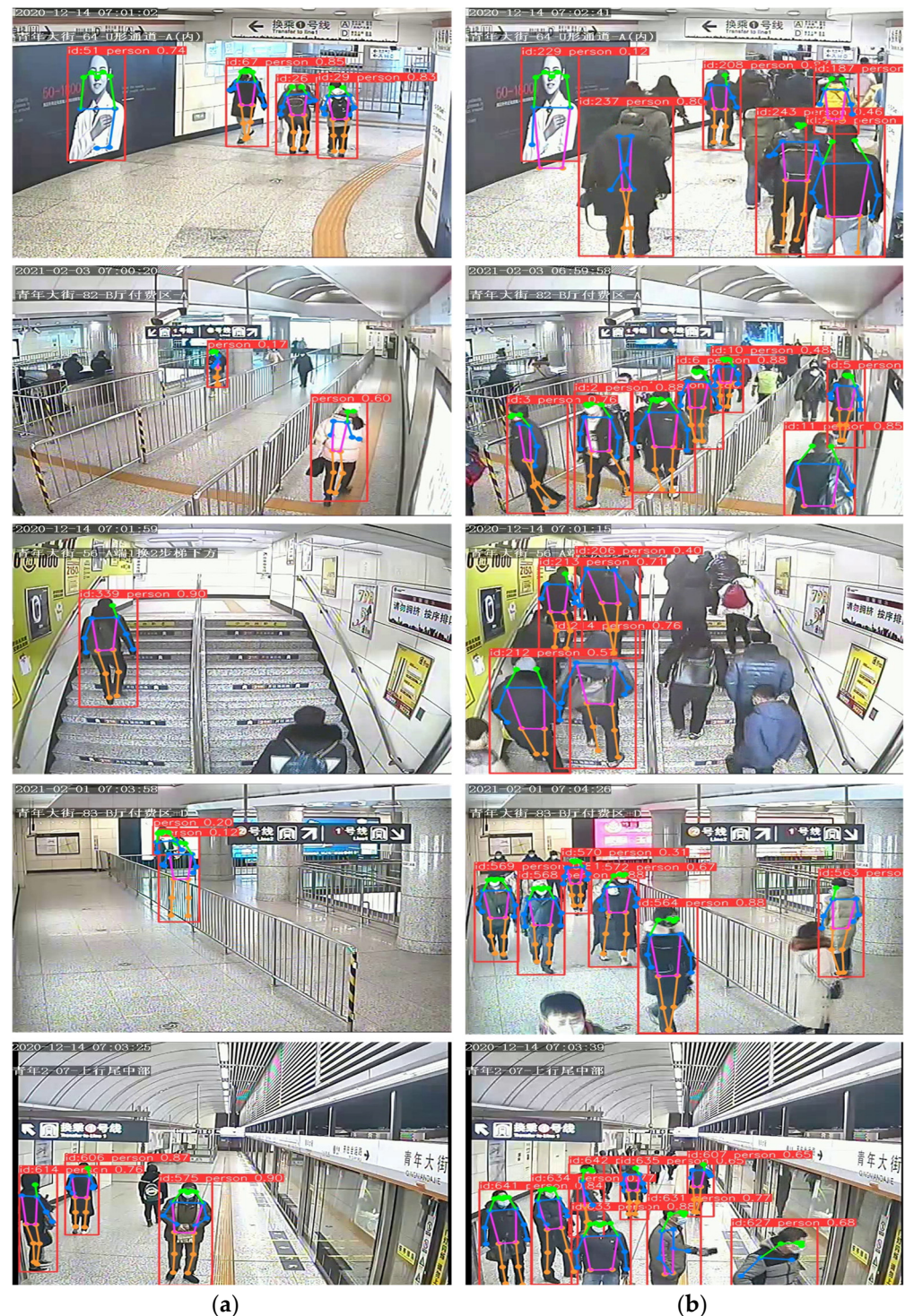


Figure 8. Visualization results of pose prediction by the baseline model in subway scenes: (a) Sparse passenger flow; (b) Dense passenger flow.

In dense crowd scenarios, the baseline model tends to suffer from keypoint confusion among overlapping individuals and severe missed detections due to occlusions. The proposed method mitigates these issues by leveraging the OrthoNets orthogonal attention mechanism to suppress feature interference between adjacent persons, together with the MFEB module, which adaptively adjusts receptive fields according to local crowd density. Consequently, the model significantly improves limb association accuracy. Even under partial occlusion, the proposed approach is still capable of inferring occluded joints based on visible cues, maintaining structurally complete and coherent pose outputs.

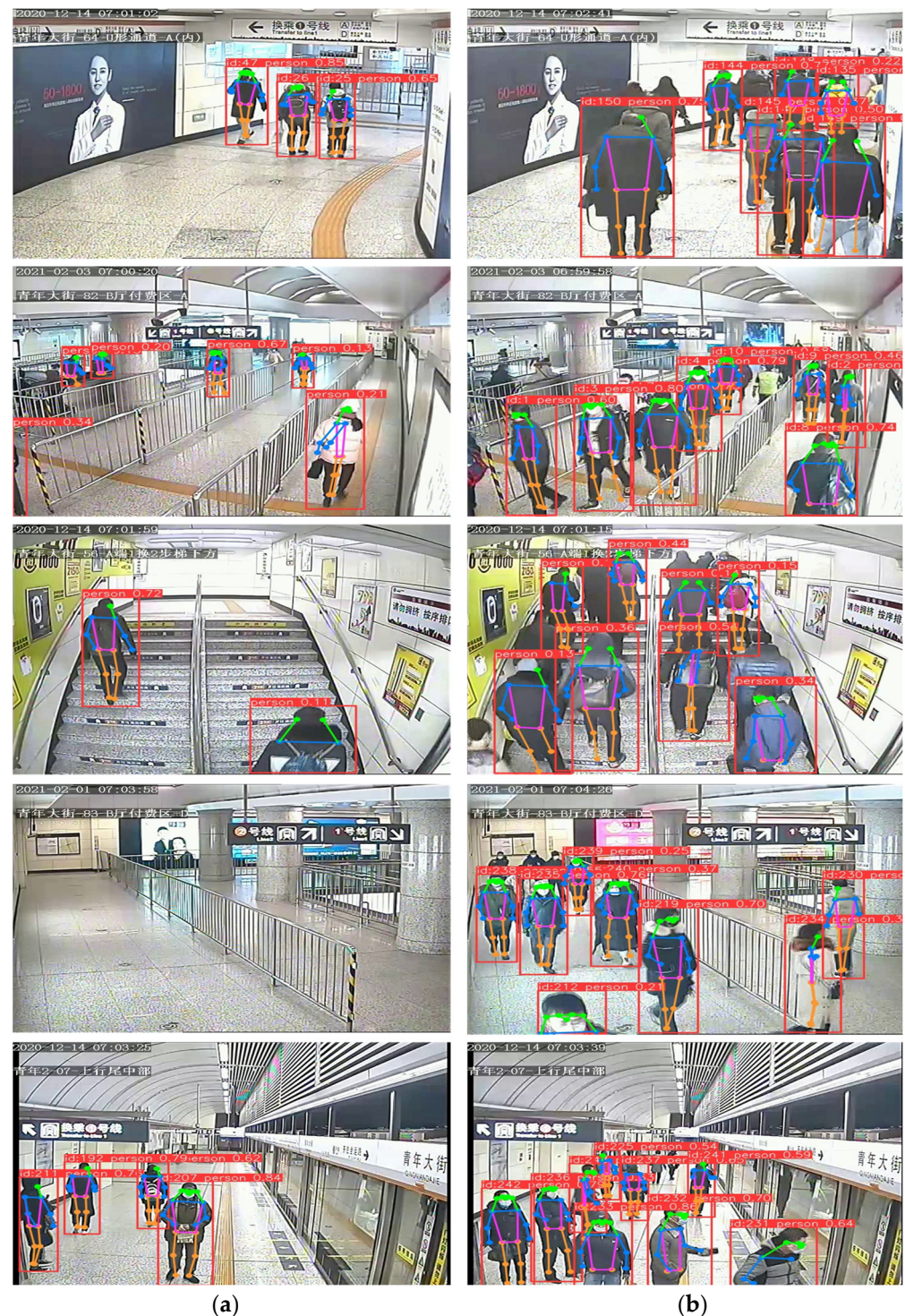


Figure 9. Visualization results of pose prediction by the improved model in subway scenes: (a) Sparse passenger flow; (b) Dense passenger flow.

For small-scale pedestrians, the proposed framework establishes a cascaded pipeline of “shallow multi-scale feature extraction–deep adaptive fusion,” which significantly enhances both bounding box quality and keypoint recall for distant targets. This effectively avoids typical failure cases observed in the baseline model, such as oversized bounding boxes covering small pedestrians or complete omission of small targets.

To further evaluate the effectiveness of the proposed method in real-world subway surveillance scenarios, we conduct an inference-only quantitative analysis on the Shenyang Metro video dataset. Since this dataset does not contain frame-level keypoint annotations

required for COCO-style mAP computation, we instead evaluate model outputs using statistical measures derived from detection and pose estimation results, which are reported in Table 4.

Table 4. Quantitative evaluation on Metro video dataset.

Metric	YOLOv8-Pose	Ours	Lightweight Ours
Avg. detection confidence	0.81	0.86	0.84
Avg. keypoint confidence	0.74	0.79	0.77
Avg. persons detected per frame	6.3	6.8	6.5
Stable tracking ratio (%)	82.1	88.4	86.7
Inference FPS	38	38	76
Latency (ms/frame)	26.3	26.3	13.2

As shown in Table 4, the proposed method consistently achieves higher detection confidence, improved keypoint reliability, and better stability compared with the baseline method, demonstrating its robustness in real-world crowded subway environments. In terms of computational efficiency, the proposed model introduces multiple cooperative modules, resulting in a 27.6% increase in parameter size compared with the baseline. Under the same hardware configuration, it achieves an inference speed of approximately 38 FPS, still satisfying real-time requirements. After lightweight deployment and half-precision optimization, the parameter size is reduced to 3.99M, while inference speed is increased to 76 FPS, representing a 68.9% improvement over the baseline, with a single-frame latency of only 13.2 ms.

Overall, these results demonstrate that the lightweight version maintains competitive accuracy while significantly improving inference efficiency, making it suitable for deployment on edge devices. The visualization results further confirm that the proposed method performs robustly in challenging scenarios involving occlusion and scale variation.

4. Conclusions

This paper presents a lightweight and efficient human pose estimation framework based on an improved YOLOv8-Pose architecture for complex subway surveillance scenarios. The proposed method integrates multi-scale feature modeling, high-frequency detail enhancement, orthogonal attention, and adaptive feature fusion to improve robustness under occlusion and scale variation.

Extensive experiments on COCO 2017 and real-world subway datasets demonstrate that the proposed method consistently improves accuracy while meeting real-time constraints. The lightweight version further reduces computational complexity with minimal performance degradation, enabling efficient deployment on edge devices. Overall, the proposed framework achieves a favorable balance between accuracy, efficiency, and robustness, making it suitable for intelligent transportation systems.

Author Contributions: Conceptualization, W.Y.; methodology, G.L.; validation, Y.Z. and Y.L.; formal analysis, Y.Z.; investigation, T.S.; resources, G.L.; data curation, G.L.; writing—original draft preparation, Y.Z.; writing—review and editing, W.Y.; project administration, T.S.; visualization, Y.L.; funding acquisition, W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (under Grant Number U23A20603).

Data Availability Statement: The original contributions of the study are included in the article, and further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
2. Ye, R.; Zhang, K.; Zhang, C.; Yan, B.; Zhou, X. Research on 3D Convolutional Autoencoder Enhanced Metro Abnormal Behavior Detection Based on Multi-Level Attentional Memory. *Multimed. Tools Appl.* **2025**, *84*, 23861–23879.
3. Zhang, Z.; Shin, S.Y. Two-Dimensional Human Pose Estimation with Deep Learning: A Review. *Appl. Sci.* **2025**, *15*, 7344. [\[CrossRef\]](#)
4. Dubey, S.; Dixit, M. A comprehensive survey on human pose estimation approaches. *Multimed. Syst.* **2023**, *29*, 167–195.
5. Nogueira, A.F.R.; Oliveira, H.P.; Teixeira, L.F. Markerless Multi-View 3D Human Pose Estimation: A Survey. *Image Vis. Comput.* **2025**, *155*, 105437. [\[CrossRef\]](#)
6. Gao, Z.; Chen, J.; Liu, Y.; Jin, Y.; Tian, D. A Systematic Survey on Human Pose Estimation: Upstream and Downstream Tasks, Approaches, Lightweight Models, and Prospects. *Artif. Intell. Rev.* **2025**, *58*, 68. [\[CrossRef\]](#)
7. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *ACM Comput. Surv.* **2024**, *56*, 11.
8. Xu, L.H.; Zhao, L.; Sun, X.X.; Yan, K.D.; Li, G.Y. A Comprehensive Review of Progress in Deep-Learning-Based Occluded Human Pose Estimation. *J. Image Graph.* **2024**, *29*, 3529–3542. [\[CrossRef\]](#)
9. Esmail, M.A.; Wang, J.; Wang, Y.; Sun, L.; Zhu, G.; Zhang, G. Resource-Aware Strategies for Real-Time Multi-Person Pose Estimation. *Image Vis. Comput.* **2025**, *146*, 105324.
10. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186.
11. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
12. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
13. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.L.; Lu, C. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7157–7173. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 19–20 June 2022; pp. 2637–2646, 2649–2658.
15. Lu, T.; Cheng, K.; Hua, X.; Qin, S. KSL-POSE: A Real-Time 2D Human Pose Estimation Method Based on Modified YOLOv8-Pose Framework. *Sensors* **2024**, *24*, 6249. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11313–11322.
17. Salman, H.; Parks, C.; Swan, M.; Gauch, J. OrthoNets: Orthogonal Channel Attention Networks. In Proceedings of the IEEE International Conference on Big Data (Big Data), Sorrento, Italy, 15–18 December 2023; pp. 3125–3132.
18. Han, Z.; Yang, H.; Wu, S.; Chen, L. Action Recognition Network Combining Spatio-Temporal Adaptive Graph Convolution and Transformer. *J. Electron. Inf. Technol.* **2024**, *46*, 2587–2595.
19. Gu, X.J.; Guo, Z.B. Dense Crowd Pose Estimation Algorithm for In-Layer Adjustment Feature Pyramid. *Semicond. Optoelectron.* **2024**, *45*, 931–938.
20. Wei, H.; Liu, X.; Xu, S.; Dai, Z.; Dai, Y.; Xu, X. DWRSeg: Dilation-Wise Residual Network for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
21. Li, A.; Zhang, L.; Liu, Y.H.; Zhu, C. Feature Modulation Transformer: Cross-Refinement of Global Representation via High-Frequency Prior for Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12480–12490.
22. Xiao, H.; Qin, J.; Jeon, S.; Yan, B.; Yang, X. MFEN: Lightweight Multi-Scale Feature Extraction Super-Resolution Network in Embedded System. *Microprocess. Microsyst.* **2022**, *93*, 104568. [\[CrossRef\]](#)
23. Parida, S.K.; Bogner, J.; Fritzsche, J. How Do Model Export Formats Impact the Development of ML-Enabled Systems? A Case Study on Model Integration. In Proceedings of the International Conference on AI Engineering—Software Engineering for AI (CAIN), Ottawa, ON, Canada, 27–28 April 2025.
24. Dorrich, M.; Fan, M.; Kist, A.M. Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks. *IEEE Access* **2023**, *11*, 57627–57634. [\[CrossRef\]](#)

25. Shafi, O.; Rai, C.; Sen, R.; Ananthanarayanan, G. Demystifying TensorRT: Characterizing Neural Network Inference Engine on Nvidia Edge Devices. In Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), Storrs, CT, USA, 7–9 November 2021; pp. 226–237.
26. Kustikova, V.; Vasiliev, E.; Khvatov, A.; Kumbrasiev, P.; Vikhrev, I.; Utkin, K.; Dudchenko, A. Intel Distribution of OpenVINO Toolkit: A Case Study of Semantic Segmentation. In *Proceedings of the Analysis of Images, Social Networks and Texts (AIST)*; Springer: Cham, Switzerland, 2019.
27. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
28. Neff, C.; Sheth, A.; Furgurson, S.; Middleton, J.; Tabkhi, H. EfficientHRNet: Efficient and Scalable High-Resolution Networks for Real-Time Multi-Person 2D Human Pose Estimation. *J. Real-Time Image Process.* **2021**, *18*, 997–1013. [[CrossRef](#)]
29. Wang, Y.H.; Li, M.Y.; Cai, H.; Chen, W.M.; Han, S. Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 13126–13136.
30. Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; Chen, K. RTMPose: Real-Time Multi-Person Pose Estimation Based on MMPose. *arXiv* **2023**, arXiv:2303.07399.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.