Received: 9 November 2025 | Revised: 6 February 2026 | Accepted: 27 February 2026 | Published online: 2 June 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–9
DOI: 10.47852/bonviewJDSIS62028142

BON VIEW PUBLISHING

A Lightweight Human Pose Estimation Algorithm Based on Improved YOLO11-Pose

Yang Gao[1], Guanglei Qiang[1] and Fujiang Yuan[1,*]

[1] *School of Computer Science and Technology, Taiyuan Normal University, China*

Abstract: As computer vision technology is applied to human pose estimation, current human pose estimation models suffer from problems such as large computational load and parameter count and slow inference speed. This paper proposes an improved lightweight human pose estimation algorithm based on YOLO11-Pose. By introducing Dysample dynamic upsampling in the Neck section to replace nearest neighbor interpolation upsampling, the accuracy of human body keypoint recognition is improved; RCSPELAN is proposed, and the C3k2 module in the model is replaced to enhance feature extraction and reduce computation and parameter requirements; meanwhile, a DESD detection head is proposed by employing detail-enhancement convolution, which effectively captures key details of the human body and reduces model parameters and computational complexity. On the MS COCO human keypoint dataset, mAP50 increased by 0.3%. At the same time, the number of model parameters decreased by 29%, and the computational cost dropped by 13.5%, resulting in a lightweight model.

Keywords: human pose estimation, Dysample, RCSPELAN, DESD, lightweight networks

1. Introduction

Human pose estimation localizes key human joints from visual inputs, serving as a foundation for numerous applications. Over the past decade, human pose estimation has developed rapidly. Early methods relied on markers or depth sensors. Recent methods use deep learning to estimate complex poses from RGB images. As these models become deeper, they require more computation. This creates problems for real-time inference and limited hardware. It also makes deployment on mobile or embedded devices difficult. As a result, lightweight human pose estimation methods have gained increasing attention.

Traditional methods of estimating the pose mainly used depth sensors or handcrafted features. For example, Hu et al. [1] introduced Point2PartVolume, a method that leveraged depth images to estimate human body volume and pose, addressing challenges such as shape completion and body segmentation through deep learning techniques applied to point clouds derived from depth data.

Chen and Chen [2] employed feature vectors trained via the K-Nearest Neighbors (KNN) classifier to recognize dance poses, demonstrating the effectiveness of carefully designed features in static scenarios. Their approach underscores the importance of feature engineering in static pose recognition, especially when combined with traditional classifiers. Zhao et al. [3] improved upon existing models by integrating object detection modules with pose estimation frameworks, utilizing datasets augmented with handcrafted annotations to enhance recognition accuracy for construction machinery. Their work highlights how feature extraction and model improvements can be tailored to specific domains using handcrafted features and traditional computer vision techniques.

Subsequent works relied on convolutional neural networks for monocular image pose estimation. Mehta et al. [4] proposed a novel markerless motion capture dataset and training strategy to improve the 3D generalization, and Tome et al. [5] proposed a unified model jointly reasoning about 2D and 3D poses. Chen and Ramanan [6] then reported that 2D estimation in combination with matching techniques could be used to improve the performance of direct 3D regression techniques, and Chen et al. [7] introduced adversarial learning to integrate geometric priors for more structure-aware predictions. Moon et al. [8] solved the problem of 3D posture estimation in multi-person scenarios by using the method of measuring distances with cameras for perception. This method can label multiple objects in a single RGB image. Comprehensive reviews by Chen et al. [9] and Wang et al. [10] summarized these developments, highlighting persistent issues such as depth ambiguity, occlusion, and lack of real-world data.

Meanwhile, new strategies gradually emerged, including deep learning architectures, transformer-based models, and physics-based modeling methods. Wang et al. [11] proposed LitePose, which was a real-time multi-person pose estimation method and is suitable for edge devices. Their approach notably reduces latency by up to five times without compromising performance, thereby advancing the applicability of pose estimation in resource-constrained environments. Zheng et al. [12] proposed PoseFormer, which was a framework centered on the Transformer and was used to establish the connection between

***Corresponding author:** Fujiang Yuan, School of Computer Science and Technology, Taiyuan Normal University, China. Email: 202325502010@stu.tynu.edu.cn

the modeling space and time in video sequences. Through the Transformer-based architecture, PoseFormer can capture the necessary associations required for accurate 3D human pose reconstruction without relying on convolutional layers. Wang et al. [13] developed TransNet, a parallel encoder architecture that enhances the encoding of spatial-temporal features, further improving 3D pose estimation performance.

Beyond purely data-driven approaches, integrating physical and kinematic modeling has shown promise. Yuan et al. [14] introduced SimPoE, which combined image-based kinematic inference with physics-based dynamics modeling to estimate 3D human motion from monocular videos. This approach underscores the importance of considering both body kinematics and physical forces to achieve more accurate and realistic pose estimations. Additionally, Gong et al. [15] proposed DiffPose, which simplifies 3D pose estimation into a reverse diffusion process, thereby addressing the uncertainties that arise in monocular 3D pose estimation under conditions of occlusion and high-level blurring.

Accurate pose estimation is dependent on the incorporation of temporal information. Liu et al. [16] employed a deep sequential network, which utilized multiple temporal cues to extract information and employed a pose-time merging and pose residual fusion module to encode the spatio-temporal context and refine keypoint detection between video frames. This multi-frame strategy demonstrates the advantage of leveraging temporal consistency to improve pose estimation.

Over the past few years, the demand for efficient and real-time high-performance human pose estimation solutions has driven research on lightweight architectures. These models aim to maintain competitive accuracy while significantly reducing computational costs, memory usage, and power consumption, which are key factors in improving the deployment of models in mobile, embedded, or edge computing applications. The following content reviews the related research that contributed to this trend and explains its relevance to the development of an improved lightweight pose estimation framework based on YOLO11.

1.1. Previous work

The pursuit of lightweight human pose estimation goals has led to many innovative achievements in network architectures, attention, and model improvements [17]. Zang et al. [18] proposed a multi-stage attention mechanism network (LMANet), which can observe postures in infrared images under challenging conditions such as different lighting conditions, demonstrating the adaptability of the attention mechanism to specific sensing modalities. Similarly, Jeon et al. [19] explored human motion assessment based on mobile devices, emphasizing the importance of compact models for real-time feedback and widespread fitness monitoring.

The attention mechanism has also been employed to enhance feature discrimination while maintaining model efficiency [20, 21]. Liu et al. [22] proposed a polarization self-attention module, which enhances keypoint localization with minimal computational overhead. To further improve the robustness of the estimation, He et al. [23] designed a multi-angle model with an unbiased decoding strategy, integrating multi-view consistency into a lightweight framework. Zhang and Zhou [24] introduced RepNet, a reparameterized regression network, which achieves a better balance between accuracy and parameter efficiency—an important step toward deployable, high-performance models.

In the context of networks with the fewest parameters, Hirschen and Avidan [25] proposed a model based on normalized flow, with only 1000 parameters, capable of detecting anomalies in posture sequences. The simplified basic architecture has also been proven to be effective; Wang et al. [26] replaced the traditional ResNet structure with a Shuffle module, reducing the computational cost and also predicting accurate coordinates. Li et al. [27] verified their VTTransPose network on standard datasets, using lightweight transformer-based modules to achieve high accuracy.

Ding et al. [28] proposed an improved YOLO-Pose model, optimizing the real-time human pose detection model, achieving higher average accuracy, fewer parameters, and faster inference speed compared to previous human pose estimation models [29]. Despite the major advances in human pose estimation in recent years, there are still some unresolved challenges. Most existing methods improve recognition accuracy by increasing model complexity or introducing complex network architectures but pay little attention to the limitations of model parameters, computational costs, and inference efficiency in practical scenarios. Therefore, many new methods, although achieving high accuracy, come at the cost of increased memory consumption and computational overhead, which limits their applicability. Therefore, enhancing the equilibrium between accuracy and efficiency necessitates further investigation.

1.2. Contributions

The problems of high computation and large model size in methods are solved by this paper's proposal of a lightweight human pose estimation algorithm based on Revised YOLO11-Pose. The method optimizes the network structure and improves key modules in the pose estimation model. It boasts equilibrium between reliability and expediency, maintaining competitive pose estimation performance. At the same time, it greatly reduces parameters and computation. The main contributions are enumerated below:

1) Dysample dynamic sampling technology is incorporated into the neck of YOLO11, which enhances the precision of human body keypoint recognition without necessitating a rise in computation and parameters.
2) RCSPELAN combines the strengths of several structures and removes redundant information from the original model.
3) A DESD detection head is proposed, and the YOLO11-Pose detection head is optimized. This method uses detail-focused convolution and shared parameters. The model captures human keypoint details more clearly and uses fewer parameters. The model also reduces computational cost.

2. Related Work

2.1. YOLO11-Pose

YOLO11-Pose is an important extension of YOLO series models in the field of human pose estimation. It innovatively integrates object detection and keypoint detection tasks into a unified end-to-end network framework. The model inherits the high-speed and high-precision characteristics of the YOLO series and is optimized for human keypoint detection, realizing real-time and accurate human pose estimation, so that the network can complete the two tasks synchronously.

Its network architecture follows the Backbone–Neck–Head design. The function of the backbone network is to extract multilevel features from imagery. The neck network fuses features

efficiently. It combines deep features with rich meaning and shallow features with clear position information. This design helps the model locate keypoints accurately. The core innovation of YOLO11-Pose lies in the design of its head network. The keypoint head is introduced in parallel on the basis of the traditional detection heads used for outputting bounding boxes and confidence. This detection head does not directly regress the absolute coordinates of keypoints but locates them by predicting the heatmap associated with each target anchor point. This method can make better use of spatial context information, especially for keypoint recognition and detection in occluded or complex scenes. This model can also simultaneously output the category confidence, the corresponding object category boxes, and the relative human keypoints of each detected human target in the image. It effectively avoids the process of the traditional two-stage method and strikes a balance between speed and accuracy.

2.2. Proposed algorithm

Regarding the high computational and numerous parameters in human pose estimation models, we propose a lightweight human pose estimation algorithm based on YOLO11-Pose. In the Neck part of YOLO11-Pose, Dysample dynamic upsampling is introduced to replace the nearest neighbor interpolation upsampling in the original model, so that the model can improve the recognition accuracy of human keypoints without increasing the amount of computation and parameters. In addition, replace the C3k2 module with RCSPELAN. It combines the advantages of CSPNet, ELAN, and RepConv to reduce parameters and computation. A DESD detection head is proposed to replace the original YOLO11-Pose detection head. It uses detail-enhancement convolution to improve the comprehensive capture of detailed information of human body keypoints, while adopting a shared parameter method to reduce parameters and computation. The revised model is illustrated in Figure 1.

2.2.1. Dysample dynamic upsampling

Traditional upsampling methods use fixed interpolation rules and cannot adaptively adjust the upsampling method according to the image content, resulting in limited detail recovery and blurred edges. Later, upsampling operators based on dynamic convolution emerged, but they still have problems such as high computational complexity in human pose estimation. This paper proposes an improved strategy that introduces Dysample [30] dynamic upsampling technology in the neck region, which improves the accuracy of human keypoint recognition without significantly increasing parameters and computation. The diagrams are depicted in Figures 2 and 3.

Giving an input feature map X of size $C \times H \times W$ and upsampling factor S, point sample set of size $sH \times sW \times 2g$. The feature is resampled by a sampling point generator to generate a high-resolution feature X' of size $sH \times sW \times C$, as shown in Equation (1).

$$X' = \text{grid\_sample}(X, S) \tag{1}$$

The generated offset O and the original sampling grid G form the sampling set of the points. The feature maps are passed through the linear layer to generate offset O of size $2s^2 \times H \times W$, and they are recombined into high-resolution original sampling grid G of size $2g \times sH \times sW$. The result of adding the offset O to the original sampling grid G is the point sampling set. The formulas

Figure 1
Model architecture

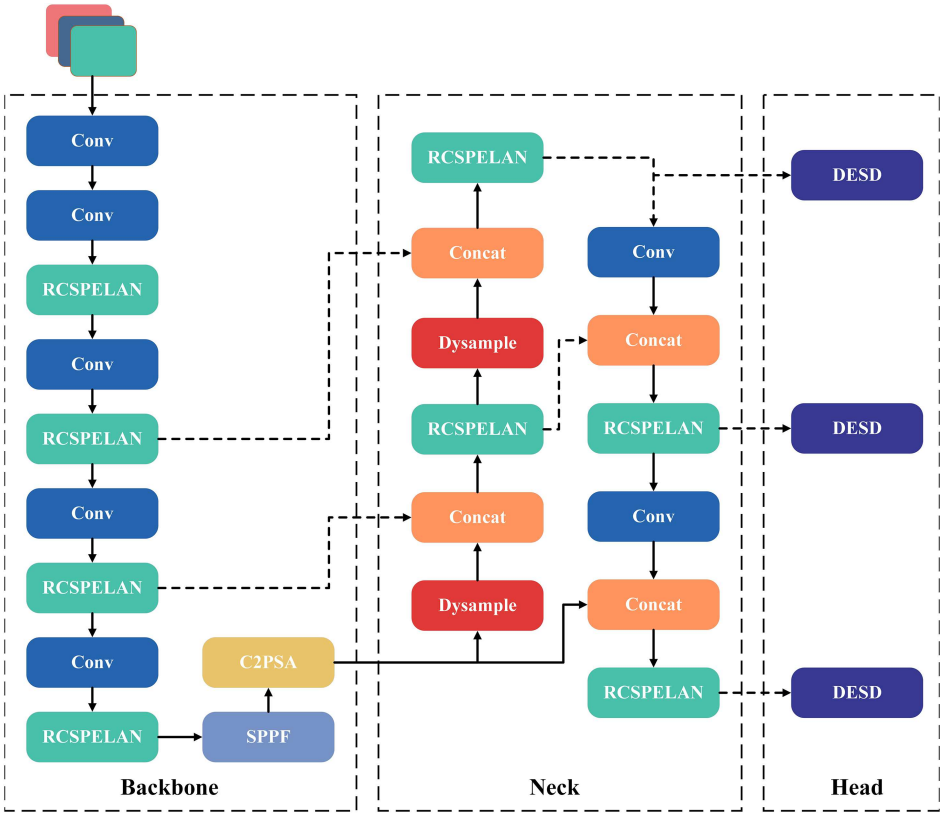


Figure 2
Dysample dynamic upsampling

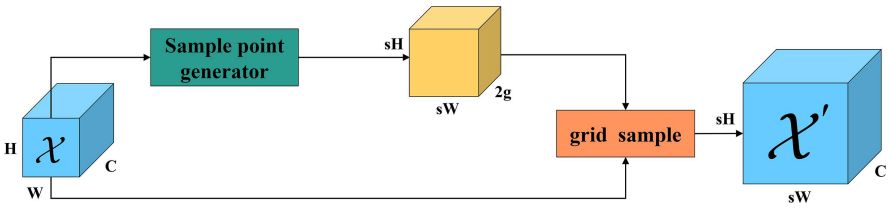
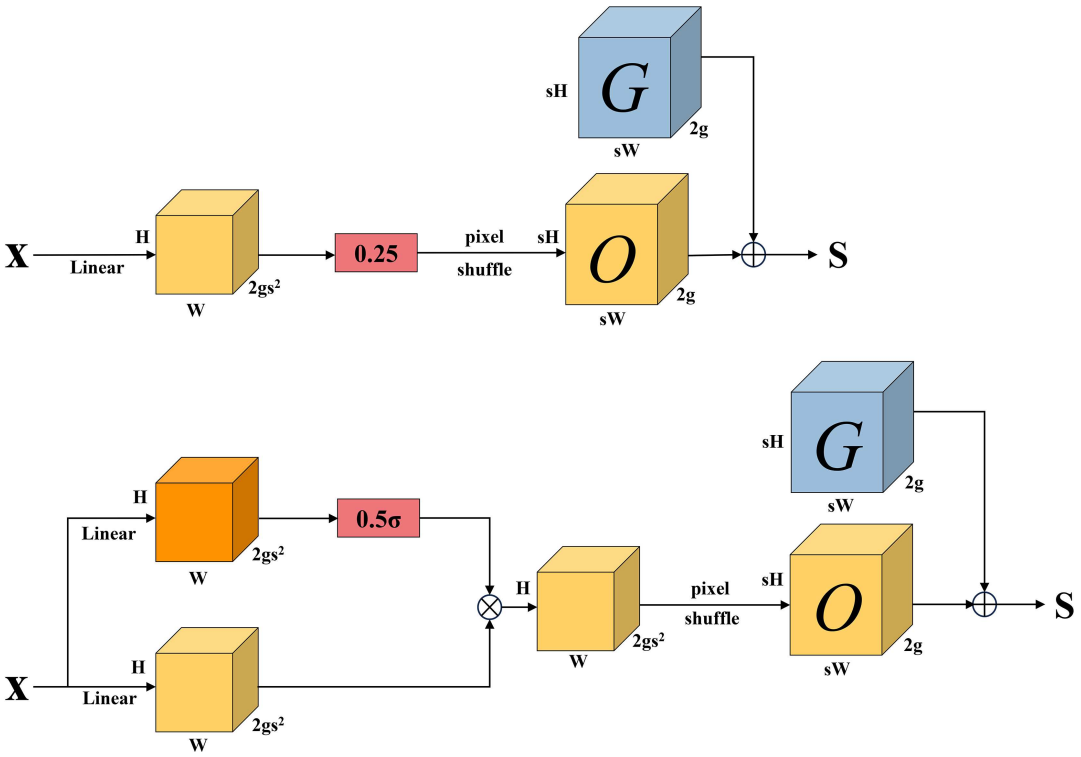


Figure 3
Point sampling generator



for the offset O and the sampling set S are given in Equations (2) and (3), respectively.

$$O = \text{linear(X)} \tag{2}$$

$$\text{S} = G + \text{O} \tag{3}$$

Dysample introduces an upsampling method that learns sampling offsets and their associated weights in an adaptive manner, tailoring them to local feature distributions. In keypoint and boundary prediction tasks, fixed upsampling methods often suffer from limited receptive-field flexibility, leading to blurred responses and localization errors near structural discontinuities. Dynamic sampling enables the model to adjust the effective receptive field, making it more focused on meaningful areas, such as human joint points or object boundary contours. The process entails the alignment of disparate features, both low-resolution and high-resolution, within the feature map. Therefore, dynamic sampling can effectively alleviate the keypoint positioning errors caused by interpolation ambiguity and improve the boundary prediction accuracy. The specific calculation formulas for dynamic sampling are shown in Equations (4) and (5).

$$O = 0.25\text{linear(X)} \tag{4}$$

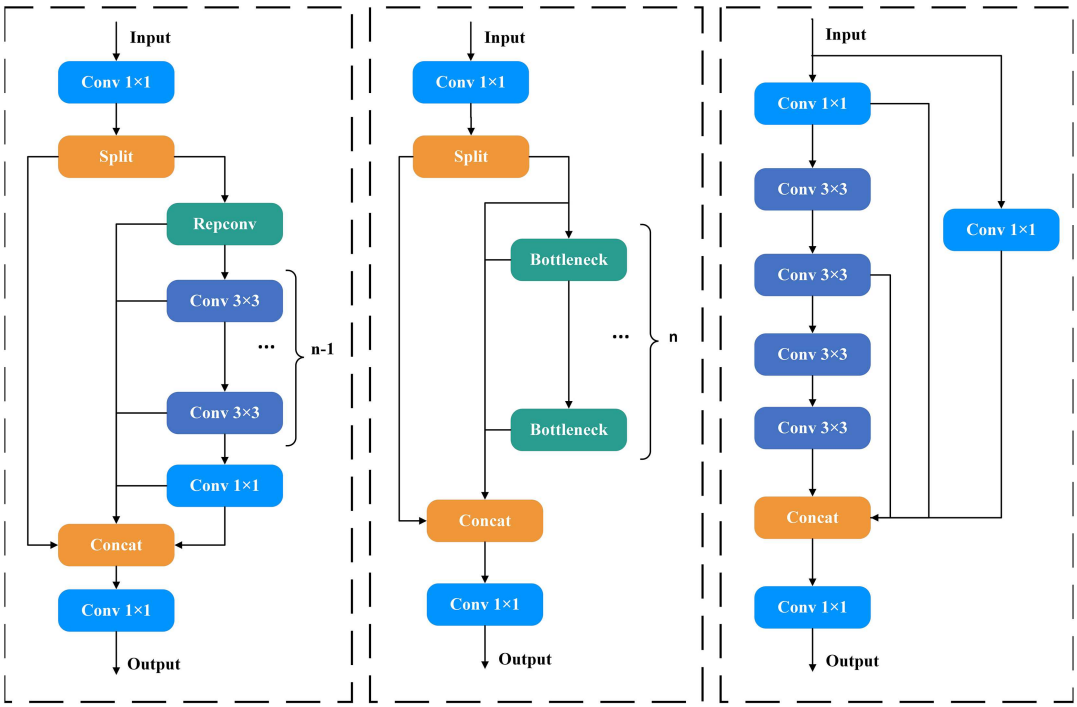
$$O = 0.5\text{sigmoid}(\text{linear}_1(\text{X})) \cdot \text{linear}_2(\text{X}) \tag{5}$$

2.2.2. RCSPELAN

Inspired by CSPNet [31] and ELAN [32], RCSPELAN integrates cross-stage partial connection and parallel path feature aggregation to optimize the effective gradient length. By dividing the path into multiple branches and allowing convolution operations and feature propagation to proceed along different paths, the proposed structure effectively alleviates gradient vanishing and gradient redundancy, while enhancing the information interaction between features at different depths.

In addition, RepConv [33] is selectively introduced into the main branch, enabling multi-branch feature modeling during the training stage while preserving a single-branch, computationally efficient structure during inference. This design achieves a balance among feature expression capability, parameters, and computation. The structure is shown in Figure 4.

Figure 4
RCSPELAN, CSPNet, and ELAN



(a) RGCSPELAN (b) CSPNet (c) ELAN

Given a feature map, the input image is first compressed using a 1 × 1 convolution. Then, the two branches are input separately. The first branch is first processed by RepConv, which allows the feature map to adopt a multi-branch structure during training and be mapped to a single branch during inference, effectively reducing the number of parameters and computation. Subsequently, $n-1$ convolutions are used to further extract features. At last, the characteristics of the two divisions are combined and presented using a 1 × 1 convolution. RCSPELAN combines the advantages of multiple architectures to reduce computation and parameter requirements.

2.2.3. DESD

The YOLO11-Pose detection head employs a decoupled structure, which improves the model's robustness, and generalization can be improved by processing semantic information at different scales, but it increases the number of parameters and loses some detailed information. A DESD detection head is proposed. This can reduce parameters. It can also enrich the edge detailed information. The specific structure is shown in Figure 5.

After generating the outputs of P3, P4, and P5, shared convolution is adopted to reduce computational cost. Although shared convolution effectively saves computational resources, the coupling of multi-scale features under shared weights inevitably limits representational capacity and leads to a certain degradation in accuracy. To address this issue, Group Normalization (GN) is first introduced to replace Batch Normalization. The mean and variance of each group of channels are computed by GN, which also divides the channels into groups, which stabilizes feature distributions and preserves inter-channel dependencies under shared weights, thereby alleviating response inconsistency caused by shared convolution. Subsequently, a deconvolution operation Deconv [34] is introduced for detail enhancement. By explicitly increasing spatial resolution and strengthening local detail representation, Deconv helps recover keypoint and edge-related features that are weakened during shared convolution. Furthermore, through the computation and fusion of five parallel convolutional branches, the capability of capturing edge information is enhanced, and deeper feature representations are provided, which further improves recognition performance.

Through the joint effect of GN and Deconv, the proposed DESD detection head effectively mitigates the accuracy degradation introduced by shared convolution while maintaining computational efficiency. Finally, a data normalization layer (Scale) is appended to the network to adaptively adjust feature scales and better account for target features at different resolutions.

3. Experiments and Analysis

3.1. Dataset

MS COCO 2017 is one of the most widely used benchmark datasets in the field of human pose estimation. It provides standardized annotations, diverse human poses, and complex real-world scenarios; therefore, performance on this dataset is generally considered to be representative.

The present document employs the MS COCO 2017 pose estimation dataset. MS COCO is a representative public dataset in the field of vision, and it is suitable for tasks requiring object detection and pose estimation. The dataset for human pose estimation defines 17 keypoints.

The dataset under scrutiny encompasses 56,599 training images and 2346 validation images. These cover various complex scenarios, including single-person and multi-person scenes, and

Figure 5
DESD

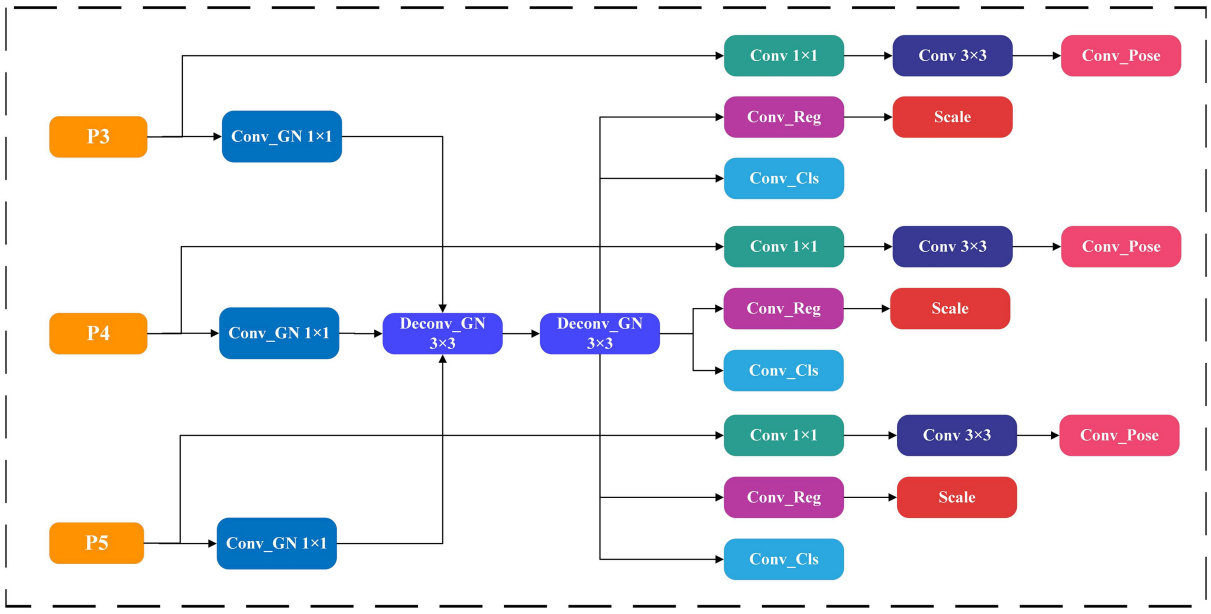


Table 1
Experimental environment

Name	Version
CPU	Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
GPU	RTX 3090 24GB
Programming language	Python3.10
Operating system	Ubuntu20.04
Deep learning framework	PyTorch2.2.0+CUDA 12.1

demonstrate diversity and complexity in human posture under different conditions.

3.2. Experimental environment

This article implements deep learning using the PyTorch framework. The experimental setup is shown in Table 1. The settings are as follows: epochs = 300, batch = 32, optimizer = SGD, learning rate = 0.01.

3.3. Evaluation indicators

Precision (P), recall rate (R), and the mean average precision (mAP50) are used as the main evaluation metrics in this paper. In the evaluation criteria for pose estimation, the Intersection over Union (IoU) of the bounding boxes is not used to determine true positives (TP) and false positives (FP). Instead, the predicted keypoints are calculated to determine whether their positions are within a reasonable error range relative to the true keypoints. The Object Keypoint Similarity (OKS) used in this work is computed, as shown in Equation (6).

$$L_{\text{oks}} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \cdot \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{6}$$

In the formula, i denotes the keypoint index; d_i^2 represents the squared Euclidean distance between the ground-truth and predicted keypoint positions; v_i indicates whether the keypoint is visible; s^2 denotes the object scale area; k_i is the standardization constant for the i-th keypoint; and δ is the indicator function, indicating that only labeled visible keypoints are considered in the calculation. Formally, precision, recall, average precision (AP), and mean average precision (mAP) are defined as in Equations (7)–(10).

$$P_{\text{kpt}} = \frac{TP_{\text{kpt}}}{TP_{\text{kpt}} + FP_{\text{kpt}}} \tag{7}$$

$$R_{\text{kpt}} = \frac{TP_{\text{kpt}}}{TP_{\text{kpt}} + FN_{\text{kpt}}} \tag{8}$$

$$AP = \int_0^1 P_{\text{kpt}} \mathrm{d}(R_{\text{kpt}}) \tag{9}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{10}$$

In the formula, TP_{kpt} indicates that the keypoint is correctly identified and the L_{oks} are greater than the set threshold, FP_{kpt}

Table 2
Ablation experiments on the MS COCO 2017 dataset

Dysample	RCSPELAN	DESD	mAP@0.5	mAP@0.5:0.95	Params/M	GFLOPs
×	×	×	0.79	0.482	2.8	7.4
√	×	×	0.797	0.492	2.8	7.4
×	√	×	0.791	0.474	2.5	7.3
×	×	√	0.793	0.478	2.3	6.5
√	√	×	0.792	0.481	2.5	7.3
√	×	√	0.797	0.48	2.3	6.5
√	√	√	0.793	0.475	2.0	6.4

indicates that a non-key point area is incorrectly identified as a keypoint and the L_{oks} are greater than the set threshold, and FN_{kpt} (pseudo-negative example) indicates that a real keypoint area is incorrectly identified as a non-key point.

3.4. Ablation experiment

The algorithm's effectiveness was verified using the following ablation experiment. Results are shown in Table 2.

Ablation experiments on the MS COCO2017 dataset validate the effectiveness of the proposed modules from both performance and efficiency perspectives. Introducing Dysample resulted in a 0.7% increase in mAP50 without increasing the model's parameters or computational load. This indicates that dynamic upsampling can effectively improve the accuracy of keypoint localization without incurring additional computational costs. Replacing the C3k2 with RCSPELAN reduces parameters by 10% while increasing mAP50 by 0.1%. This improvement mainly lies in the combined use of the dual-branch architecture and RepConv, eliminating redundant computations while retaining the feature fusion capability. In contrast, the DESD detection head reduces 18% of parameters and 12% of computational costs while achieving a 0.3% increase in mAP50. This is because DESD employs group normalization, Deconv, and parallel convolution structures, enhancing the feature representation related to joint positioning and edge details.

When Dysample is combined with RCSPELAN and DESD, the model's mAP50 increases by 0.2% and 0.7%, respectively, while parameters and computation are both decreased. This demonstrates the strong complementarity of dynamic upsampling and structural optimization in the human pose estimation task. Finally, integrating all three modules reduces parameters by 29% and computation by 13.5% while maintaining stability, achieving an effective balance between model lightweighting and performance improvement. It is worth noting that although the proposed model achieved a 0.3% improvement in mAP@0.5, the value of mAP@0.5:0.95 decreased by 0.7% compared to the baseline. This is normal in lightweight models. Reducing model capacity and computational complexity may have an impact on high IoU localization accuracy. Since mAP@0.5 better reflects the accuracy of keypoint detection in real scenarios, this paper prioritizes the comparison in mAP@0.5 and also accepts maintaining certain performance on more stringent evaluation metrics.

Since this paper focuses on lightweight models, GFLOPS and parameter count are also evaluation metrics. Parameter quantity represents the learnable parameters in the model, which is a system for measuring the storage and memory size of the model. Fewer parameters and lower GFLOPS values generally mean an efficient and lightweight model, better suited for deployment on resource-constrained devices.

3.5. Comparative experiment

To verify the performance of the improved algorithm presented in this paper, a comparative experiment was conducted on the MS COCO 2017 dataset, as shown in Table 3.

Through model comparison experiments, the parameters, computational load, and accuracy of YOLOv7-W6-Pose are all the highest. It is evident that this model is ill-suited for practical applications; in comparison to YOLOx-Pose-tiny, the proposed model has comparable accuracy, but its parameters are approximately one-third of that model; compared to YOLOv8n-Pose, mAP50 is 0.3% higher, and its parameters and computational load are both lower; compared with YOLO11s-Pose, although it does not have an advantage in accuracy, it has approximately one-quarter of the parameters and computational load of it.

This means that the model has fewer parameters and is less complicated to run, but it is still accurate enough to be used. This makes suited for devices with limited computing resources. Previous studies have shown that under resource-constrained

Table 3
Comparative experiment

Model	Image size/pixel	mAP@0.5	mAP@0.5:0.95	Params/M	GFLOPs
YOLOv5s6-Pose	640 × 640	0.842	0.575	15.0	20.3
YOLOx-Pose-tiny	416 × 416	0.791	0.526	6.0	4.4
YOLOv7-W6-Pose	960 × 960	0.94	0.74	80.0	101.6
YOLOv8n-Pose	640 × 640	0.79	0.489	3.2	9.2
YOLO11s-Pose	640 × 640	0.857	0.583	9.9	23.1
Our method	640 × 640	0.793	0.475	2.0	6.4

conditions, model performance is often affected by limitations in model complexity and feature representation capacity [35]. The comparative experimental results indicate that the proposed method achieves a favorable balance between computational efficiency and performance stability.

4. Conclusion

To overcome the issues of expensive computation, sizable models, and slow execution in human pose estimation models, this paper puts forward an enhanced lightweight human pose estimation algorithm that builds on YOLO11-Pose. Experimental results show that the model keeps competitive accuracy. It reduces parameters and computation effectively. This demonstrates its potential for deployment on devices with limited computing power. Because of lower computational cost, the model can be deployed on edge or mobile devices. Future work will evaluate its inference speed and latency.

Subsequent research endeavors will center on the extension of the framework to encompass real-time pose tracking for multiple individuals in complex scenarios. The goal is to improve robustness in heavy occlusion and crowded conditions while keeping real-time performance. In addition, we will explore lightweight approaches for monocular 3D human pose estimation. Additionally, the model will be assessed using larger datasets. This will enhance its generalization capabilities and prediction accuracy while maintaining a balance between performance and efficiency.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

This paper utilizes the MS COCO human keypoint dataset, which is publicly available at https://cocodataset.org/#download.

Author Contribution Statement

Yang Gao: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Guanglei Qiang:** Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Visualization, Supervision, Project administration. **Fujiang Yuan:** Methodology, Software, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision.

References

[1] Hu, P., Dai, X., Zhao, R., Wang, H., Ma, Y., & Munteanu, A. (2023). Point2PartVolume: Human body volume estimation from a single depth image. *IEEE Transactions on Instrumentation and Measurement*, *72*, 5502812. https://doi.org/10.1109/TIM.2023.3284948

[2] Chen, J., & Chen, L. (2022). Movement evaluation algorithm-based form tracking technology and optimal control of limbs for dancers. *Mathematical Problems in Engineering*, *2022*(1), 7749324. https://doi.org/10.1155/2022/7749324

[3] Zhao, J., Cao, Y., & Xiang, Y. (2024). Pose estimation method for construction machine based on improved AlphaPose model. *Engineering, Construction and Architectural Management*, *31*(3), 976–996. https://doi.org/10.1108/ECAM-05-2022-0476

[4] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017). Monocular 3D human pose estimation in the wild using improved CNN supervision. In *2017 International Conference on 3D Vision*, 506–516. https://doi.org/10.1109/3DV.2017.00064

[5] Tome, D., Russell, C., & Agapito, L. (2017). Lifting from the deep: Convolutional 3D pose estimation from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 5689–5698. https://doi.org/10.1109/CVPR.2017.603

[6] Chen, C.-H., & Ramanan, D. (2017). 3D human pose estimation = 2D pose estimation + matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 5759–5767. https://doi.org/10.1109/CVPR.2017.610

[7] Chen, Y., Shen, C., Wei, X.-S., Liu, L., & Yang, J. (2017). Adversarial PoseNet: A structure-aware convolutional network for human pose estimation. In *2017 IEEE International Conference on Computer Vision*, 1221–1230. https://doi.org/10.1109/ICCV.2017.137

[8] Moon, G., Chang, J. Y., & Lee, K. M. (2019). Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *2019 IEEE/CVF International Conference on Computer Vision*, 10132–10141. https://doi.org/10.1109/ICCV.2019.01023

[9] Chen, Y., Tian, Y., & He, M. (2020). Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, *192*, 102897. https://doi.org/10.1016/j.cviu.2019.102897

[10] Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., & Shao, L. (2021). Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, *210*, 103225. https://doi.org/10.1016/j.cviu.2021.103225

[11] Wang, Y., Li, M., Cai, H., Chen, W., & Han, S. (2022). Lite pose: Efficient architecture design for 2D human pose estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13116–13126. https://doi.org/10.1109/CVPR52688.2022.01278

[12] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., & Ding, Z. (2021). 3D human pose estimation with spatial and temporal transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, 11636–11645. https://doi.org/10.1109/ICCV48922.2021.01145

[13] Wang, C., Xiong, Z., Li, Y., Luo, Y., & Cao, Y. (2023). TransNet: Parallel encoder architecture for human pose estimation. *Smart Health*, *28*, 100395. https://doi.org/10.1016/j.smhl.2023.100395

[14] Yuan, Y., Wei, S.-E., Simon, T., Kitani, K., & Saragih, J. (2021). SimPoE: Simulated character control for 3D human pose estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7155–7165. https://doi.org/10.1109/CVPR46437.2021.00708

[15] Gong, J., Foo, L. G., Fan, Z., Ke, Q., Rahmani, H., & Liu, J. (2023). DiffPose: Toward more reliable 3D pose estimation. In *2023 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, 13041–13051. https://doi.org/10.1109/CVPR52729.2023.01253

[16] Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., & Wang, X. (2021). Deep dual consecutive network for human pose estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 525–534. https://doi.org/10.1109/CVPR46437.2021.00059

[17] Ye, C., Guo, Z., Gao, Y., Yuan, F., Wu, M., & Xiao, K. (2025). Deep learning-driven body-sensing game action recognition: A research on human detection methods based on MediaPipe and YOLO. In *2025 6th International Conference on Computer Engineering and Application*, 2087–2092. https://doi.org/10.1109/ICCEA65460.2025.11103172

[18] Zang, Y., Fan, C., Zheng, Z., & Yang, D. (2021). Pose estimation at night in infrared images using a lightweight multi-stage attention network. *Signal, Image and Video Processing*, *15*(8), 1757–1765. https://doi.org/10.1007/s11760-021-01916-3

[19] Jeon, H., Kim, D., & Kim, J. (2021). Human motion assessment on mobile devices. In *2021 International Conference on Information and Communication Technology Convergence*, 1655–1658. https://doi.org/10.1109/ICTC52510.2021.9621114

[20] Sun, Z., & Mariano, V. Y. (2025). SiT-YOLOv9: An efficient algorithm for learning behavior detection in the home environment. *Journal of Computational and Cognitive Engineering*, *4*(2), 173–185. https://doi.org/10.47852/bonviewJCCE42023949

[21] Le, A. D., Pham, D. A., Pham, D. T., & Vo, H. B. (2025). AlertTrap: A study on object detection in remote insect trap monitoring system using on the edge deep learning platform. *Journal of Computational and Cognitive Engineering*, *4*(3), 284–295. https://doi.org/10.47852/bonviewJCCE42023264

[22] Liu, S., He, N., Wang, C., Yu, H., & Han, W. (2023). Lightweight human pose estimation algorithm based on polarized self-attention. *Multimedia Systems*, *29*(1), 197–210. https://doi.org/10.1007/s00530-022-00981-z

[23] He, J., Zhang, W., Shang, R., Feng, J., & Jiao, L. (2023). Multi-angle models and lightweight unbiased decoding-based algorithm for human pose estimation. *International Journal of Pattern Recognition and Artificial Intelligence*, *37*(08), 2356014. https://doi.org/10.1142/S0218001423560141

[24] Zhang, X., & Zhou, Q. (2023). RepNet: A lightweight human pose regression network based on re-parameterization. *Applied Sciences*, *13*(16), 9475. https://doi.org/10.3390/app13169475

[25] Hirschorn, O., & Avidan, S. (2023). Normalizing flows for human pose anomaly detection. In *2023 IEEE/CVF International Conference on Computer Vision*, 13499–13508. https://doi.org/10.1109/ICCV51070.2023.01246

[26] Wang, Z., Zheng, R., & Chen, Y. (2024). Lightweight 2D human pose estimation based on simple coordinate classification. In *Fifteenth International Conference on Graphics and Image Processing: Proceedings of SPIE*, *13089*, 130890L. https://doi.org/10.1117/12.3020930

[27] Li, R., Li, Q., Yang, S., Zeng, X., & Yan, A. (2024). An efficient and accurate 2D human pose estimation method using VTTransPose network. *Scientific Reports*, *14*(1), 7608. https://doi.org/10.1038/s41598-024-58175-8

[28] Ding, J., Niu, S., Nie, Z., & Zhu, W. (2024). Research on human posture estimation algorithm based on YOLO-Pose. *Sensors*, *24*(10), 3036. https://doi.org/10.3390/s24103036

[29] Yuan, F., Huang, X., Jiang, H., Jiang, Y., Zuo, Z., Wang, L., . . . , & Peng, Y. (2025). An xLSTM-XGBoost ensemble model for forecasting non-stationary and highly volatile gasoline price. *Computers*, *14*(7), 256. https://doi.org/10.3390/computers14070256

[30] Liu, W., Lu, H., Fu, H., & Cao, Z. (2023). Learning to upsample by learning to sample. In *2023 IEEE/CVF International Conference on Computer Vision*, 6004–6014. https://doi.org/10.1109/ICCV51070.2023.00554

[31] Wang, C.-Y., Mark Liao, H.-Y., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1571–1580. https://doi.org/10.1109/CVPRW50498.2020.00203

[32] Wang, C.-Y., Mark Liao, H.-Y., & Yeh, I.-H. (2023). Designing network design strategies through gradient path analysis. *Journal of Information Science and Engineering*, *39*(4), 975–995. https://doi.org/10.6688/JISE.202307_39(4).0016

[33] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). RepVGG: Making VGG-style convnets great again. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13728–13737. https://doi.org/10.1109/CVPR46437.2021.01352

[34] Chen, Z., He, Z., & Lu, Z.-M. (2024). DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, *33*, 1002–1015. https://doi.org/10.1109/TIP.2024.3354108

[35] Zhao, H., Yan, L., Hou, Z., Lin, J., Zhao, Y., Ji, Z., & Wang, Y. (2025). Error analysis strategy for long-term correlated network systems: Generalized nonlinear stochastic processes and dual-layer filtering architecture. *IEEE Internet of Things Journal*, *12*(16), 33731–33745. https://doi.org/10.1109/JIOT.2025.3578285

How to Cite: Gao, Y., Qiang, G., & Yuan, F. (2026). A Lightweight Human Pose Estimation Algorithm Based on Improved YOLO11-Pose. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62028142