Improving Human Keypoint Detection with Optimized YOLOv11

Hongjiang Li
Northeast Petroleum University, Daqing 163318, Heilongjiang, China

Abstract: To address the challenges of inaccurate localization of multi-scale human keypoints and keypoint drift under dynamic pose variations, this study proposes an efficient keypoint detection approach built upon an improved YOLOv11 architecture. A Multi-Resolution Parallel Network (MRPN) is introduced after the backbone to maintain parallel processing of high-, medium-, and low-resolution feature maps while enabling cross-scale interaction, thereby enhancing the model's capability to perceive keypoints of varying scales. In addition, a Temporal Spatial Context Fusion Module (TSCFM) is designed, which integrates a dynamic temporal modeling unit with a spatial context enhancement layer to jointly optimize spatial structural consistency and temporal coherence across both single-frame and sequential-frame inputs. The proposed method significantly improves the accuracy and robustness of keypoint detection while preserving the inference efficiency of the original YOLOv11, making it well suited for real-time applications requiring high precision.

Keywords: Human keypoint detection; YOLO; multi-resolution feature fusion; real-time pose estimation.

1. Introduction

Human keypoint detection, as one of the core tasks in computer vision, plays a pivotal role in a wide range of applications, including action recognition, human-computer interaction, sports analytics, virtual reality, and medical rehabilitation [1]. With the rapid development of computer-vision technology and the widespread use of intelligent devices, keypoint detection has become an essential bridge connecting the physical and digital worlds, enabling various intelligent systems to perceive, interpret, and respond to human motion. In sports training, for example, it supports the analysis of motion quality; in rehabilitation, it helps evaluate the recovery of motor functions; and in human-computer interaction, it facilitates more natural and intuitive body-motion control [2].

Despite its broad applicability, human keypoint detection remains a challenging task. One major difficulty arises from the large variation in keypoint scales: small landmarks such as facial features coexist with larger joints such as shoulders and hips, making it difficult for traditional feature extractors to capture discriminative representations across all scales. Moreover, human poses are highly dynamic, and the positions of keypoints can change rapidly across consecutive frames, leading to temporal instability and localization drift[3]. Furthermore, explicit geometric relationships exist among keypoints — such as joint angles and relative spatial constraints — which are essential for producing structurally coherent and anatomically plausible predictions but are often insufficiently modeled by existing methods.

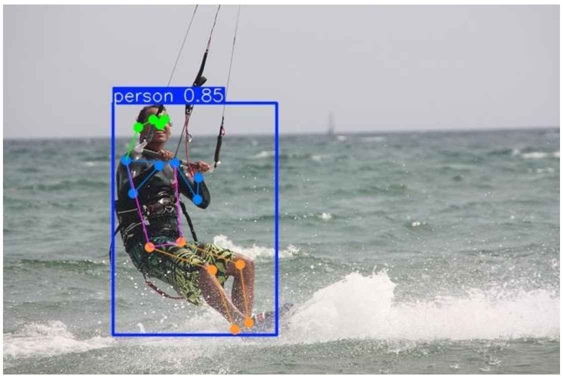

Figure 1. Keypoint Detection Figure

The effect diagrams are shown in Figures 1. Deep-learning-based approaches, particularly those inspired by the YOLO family, have recently become the mainstream solution for real-time detection tasks due to their high efficiency. YOLOv11, as a representative single-stage detector, consists of a backbone network for feature extraction, a feature pyramid for multi-scale fusion, and a detection head for predicting keypoint locations [4]. However, the standard framework exhibits limitations when directly applied to keypoint detection: it struggles to handle multi-scale keypoints effectively, especially fine-grained landmarks; it lacks mechanisms to maintain temporal coherence during rapid pose changes; and it does not explicitly model geometric relationships among keypoints, which may result in degraded accuracy in complex or highly deformable poses.

To address these issues, this paper proposes a real-time human keypoint detection model based on an improved YOLOv11 architecture [5]. The proposed approach incorporates a Multi-Resolution Parallel Network (MRPN) to better capture keypoints of varying scales through parallel feature processing and cross-resolution fusion [6]. Additionally, a Temporal Spatial Context Fusion Module (TSCFM) is designed to enhance temporal continuity and structural consistency by integrating dynamic temporal modeling with spatial contextual reasoning [7]. Through these improvements, the model aims to achieve robust and accurate keypoint localization while maintaining the efficiency characteristic of the YOLO series.

The remainder of this paper is organized as follows. Chapter 2 reviews existing research on human keypoint detection. Chapter 3 introduces the architecture and key components of the proposed approach, including the MRPN and TSCFM modules. Chapter 4 presents the experimental setup and evaluation methodology. Chapter 5 concludes the work and discusses potential future research directions.

2. Related Work

Human keypoint detection, as a fundamental task in computer vision, has a research history dating back to the 1990s [8]. Early approaches relied heavily on hand-crafted features and graph-based models, leveraging prior knowledge of human anatomy — such as joint-length proportions and feasible angle ranges — to localize keypoints. However, these methods exhibited poor adaptability to complex scenarios, including occlusion, extreme poses, and challenging illumination conditions [9].

With the rise of deep learning, keypoint detection methods have undergone substantial evolution. A major milestone appeared in 2016 with the introduction of the Stacked Hourglass Network, a heatmap-regression－based architecture that achieved high accuracy by stacking multiple prediction modules [10]. Despite its effectiveness, the method required high computational cost, limiting its applicability in real-time systems. Subsequently, OpenPose incorporated a body-part affinity representation and a multi-stage heatmap refinement strategy, enabling multi-person keypoint detection and becoming one of the most influential real-time frameworks at the time. Nevertheless, its heavy computation remained a barrier to deployment on mobile devices.

A significant breakthrough was marked in 2021 by HRNet, which maintained high-resolution feature representations through parallel multi-scale processing [11]. By avoiding the loss of fine-grained information typically introduced by repeated downsampling in feature pyramid networks, HRNet achieved substantial gains in accuracy and established itself as a landmark model in the field. However, HRNet primarily focuses on single-frame pose estimation and lacks mechanisms for ensuring temporal consistency under dynamic pose changes.

In 2018, YOLOv3 integrated keypoint detection into the YOLO detection pipeline, leveraging the efficiency of single-stage detectors to achieve real-time inference. While the model demonstrated favorable speed, it still struggled with multi-scale keypoints, particularly small landmarks such as fine facial features [12].

More recently, increasing attention has been directed toward modeling geometric relationships among keypoints. Part-aware Network enhanced spatial context by identifying body parts as intermediate semantic units, significantly improving detection performance in scenarios involving large joint-angle variations. Methods such as SimpleBaseline further simplified the detection pipeline while maintaining high accuracy and efficiency, delivering real-time performance without sacrificing precision.

For mobile applications, lightweight models such as MobileNet-Pose have integrated compact backbone networks with keypoint detection heads, achieving a practical balance between accuracy and computational efficiency [13]. These approaches can reach real-time speeds on resource-constrained devices, making them suitable for mobile and embedded deployments.

Overall, human keypoint detection has evolved from early hand-crafted pipelines to end-to-end deep learning models, from single-frame estimation to leveraging multi-frame information, and from accuracy-focused designs to architectures that balance precision and efficiency. Despite the progress, challenges remain in multi-scale feature processing, maintaining coherence under rapid pose changes, and effectively modeling spatial relationships among keypoints. In particular, achieving high accuracy on small-scale keypoints while preserving real-time performance, and mitigating localization drift during dynamic pose transitions, continue to be open problems — providing the motivation and innovation direction for this study.

3. Method

3.1 Multi-Resolution Parallel Network (MRPN)

One of the central challenges in human keypoint detection lies in the large variation of keypoint scales. When keypoints span a wide range of sizes—from small landmarks such as the head or eyes to larger joints such as shoulders or legs — traditional Feature Pyramid Networks (FPN) often struggle to capture sufficient multi-scale representations [14]. To address this issue, we introduce the Multi-Resolution Parallel Network (MRPN), a feature-fusion module inspired by the principle of maintaining multi-resolution features in parallel, similar in spirit to HRNet but redesigned with a distinct architecture tailored for YOLOv11.

MRPN is inserted after the YOLOv11 backbone and is motivated by the need to preserve high-resolution representations while enabling efficient multi-scale feature interaction. Unlike conventional FPNs that repeatedly downsample and upsample features, MRPN retains four parallel feature streams at different resolutions (1/4, 1/8, 1/16, and 1/32 of the original input). These feature maps remain active throughout the processing pipeline, preventing information loss caused by aggressive downsampling operations.

A key innovation of MRPN lies in its cross-scale interaction mechanism, which allows features at different resolutions to complement one another. At each resolution level, information from all other scales is integrated through newly designed cross-connections. Instead of simple concatenation, MRPN employs lightweight convolutional layers and attention-based fusion blocks to adaptively aggregate cross-scale information, enabling high-resolution features to gain global context while low-resolution features benefit from fine-grained spatial details.

During the final fusion stage, MRPN applies a feature rescaling strategy that upsamples high-resolution streams to align with low-resolution ones. This ensures consistent spatial alignment and avoids distortions often seen in traditional fusion schemes. The resulting feature representation captures both local structural detail and broader contextual cues, making it highly suitable for precise keypoint localization.

In terms of effectiveness, the MRPN module significantly enhances detection performance for both small- and large-scale keypoints. Small landmarks such as facial keypoints benefit from improved high-resolution retention, while larger joints gain more structurally consistent representations through cross-scale context modeling. The parallel architecture also reduces unnecessary computations, increasing overall efficiency. The lightweight cross-scale connections increase computational cost by only a small margin, making the module suitable for real-time and mobile-device deployment. In occlusion-heavy scenarios, MRPN demonstrates robust feature extraction capability, accurately localizing partially occluded joints and providing high-quality representations for subsequent attention modules.

Compared with HRNet, MRPN strengthens cross-scale interaction with a more efficient and lightweight design, making it well-suited for resource-constrained applications requiring both accuracy and real-time performance. The structural schematic diagram of MRPN is shown in Figure 2.

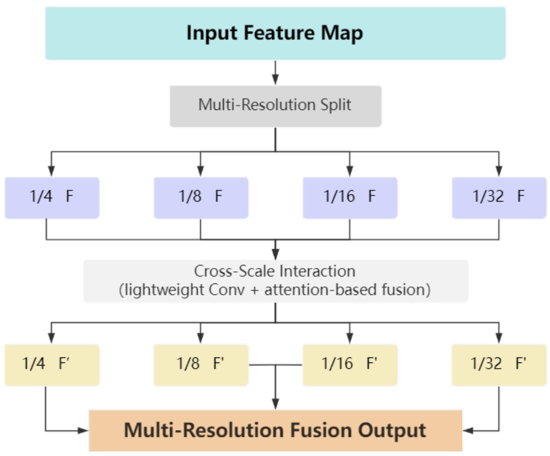

Figure 2. MRPN

3.2 Temporal Spatial Contextual Fusion Module (TSCFM)

To address the issue of keypoint localization drift during rapid pose transitions, we propose the Temporal Spatial Contextual Fusion Module (TSCFM). This module jointly models temporal continuity across frames and the spatial contextual information surrounding each joint, enabling stable and accurate keypoint estimation in dynamic scenarios [15].

TSCFM consists of two synergistic components: the Dynamic Temporal Modeling Unit (DTMU) and the Spatial Context Enhancement Layer (SCEL). Together, they refine the feature representation of each keypoint from temporal and spatial perspectives.

(1) Dynamic Temporal Modeling Unit (DTMU)

DTMU captures cross-frame motion cues by computing keypoint displacement vectors and applying an adaptive temporal filtering strategy. Its core lies in a dynamic temporal weighting mechanism that adjusts the filtering strength according to the magnitude of pose variation. During smooth movements, DTMU preserves more historical information; during intense actions such as jumping or rapid rotation, it suppresses outdated context to prevent drift.

(2) Spatial Context Enhancement Layer (SCEL)

SCEL focuses on enhancing local spatial context by employing joint-aware convolutional kernels that emphasize structural cues around each keypoint. This design is particularly beneficial for joints with large deformation or self-occlusion—such as elbows and knees—where fine-grained spatial features are crucial.

The innovation behind TSCFM lies in its unified modeling of temporal dynamics and spatial structure. Human pose is inherently a spatiotemporal signal; the keypoint position in the current frame is influenced not only by visual evidence but also by the posture evolution from previous frames. Leveraging this property, DTMU captures the underlying motion trend while SCEL reinforces local structural cues, resulting in keypoint predictions that are both stable and precise. The structural schematic diagram of TSCFM is shown in Figure 3.

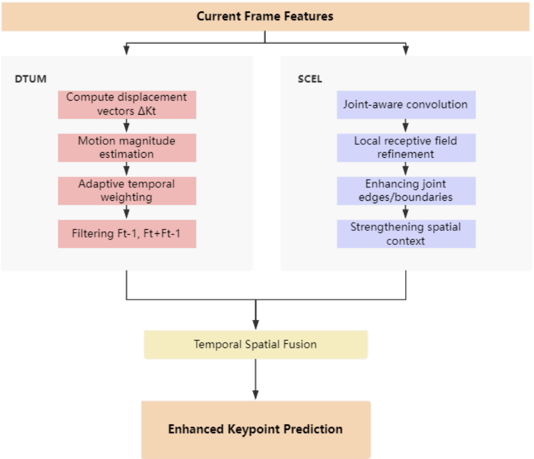
Current Frame Features
DTUM
Compute displacement vectors ΔKt
Motion magnitude estimation
Adaptive temporal weighting
Filtering Ft-1, Ft+Ft-1
SCEL
Joint-aware convolution
Local receptive field refinement
Enhancing joint edges/boundaries
Strengthening spatial context
Temporal Spatial Fusion
Enhanced Keypoint Prediction

Figure 3. TSCFM

4. Experimental Analysis

4.1 Experimental Settings

All experiments are conducted on the COCO 2017 keypoint validation set, which contains 5,000 images annotated with 17 human keypoints. To ensure consistent evaluation, all model variants—including the baseline, MRPN-enhanced model, TSCFM-enhanced model, and the final optimized YOLOv11—are trained and tested under the same experimental configuration. The input resolution is fixed at 640×640 for both training and inference. Standard data augmentation strategies commonly used in keypoint detection, such as random horizontal flipping, scale jittering, and color perturbation, are applied to improve generalization performance. Training is performed using the AdamW optimizer with an initial learning rate of 0.001 and a batch size of 32. The total training schedule spans 300 epochs, and all experiments are conducted on a single NVIDIA RTX 3090 GPU with mixed-precision acceleration enabled.

Inference speed (FPS) is measured using a batch size of 1 with the same input resolution to ensure comparability across models. The reported FPS values represent the average performance after a brief warm-up phase, thereby reflecting realistic single-image inference efficiency. No test-time augmentation or post-processing refinements beyond standard decoding are applied, and all models are evaluated strictly following the COCO keypoint metric based on OKS. Unless otherwise stated, all implementation details remain consistent across experiments to ensure that performance differences can be directly attributed to the introduction of MRPN and TSCFM.

4.2 Performance Comparison

Table 1. Performance comparison of different model variants

Model	Params (M)	FPS	AP@0.5:0.95	AP50
Baseline	22.5	52	54.2%	76.8%
+ MRPN	23.1	50	56.7%	78.5%
+ TSCFM	23.3	49	57.8%	79.2%
Optimized YOLOv11	23.5	48	58.7%	79.8%

As shown in Table 1, the optimized YOLOv11 achieve 58.7% AP@0.5:0.95, improving by 4.5 percentage points over the baseline, while maintaining real-time performance at 48 FPS. The integration of the MRPN module increases the model size by only 0.6M to 23.1M parameters, with a slight decrease of 2 FPS. However, AP@0.5:0.95 increases by 2.5 points to 56.7%, and AP50 increases by 1.7 points to 78.5%. Introducing the TSCFM module further adds 0.2M parameters and reduces FPS by 1, but brings an additional 3.6-point improvement in AP@0.5:0.95 and 2.4-point improvement in AP50. When both MRPN and TSCFM are applied together, the model size increases modestly to 23.5M, while AP@0.5:0.95 improves by 4.5 points to 58.7% and AP50 improves by 3.0 points to 79.8%, demonstrating strong combined effectiveness. The figure 4 illustrates the performance improvements brought by the proposed modules, where both AP@0.5:0.95 and AP50 consistently increase as MRPN and TSCFM are progressively integrated into the baseline model.

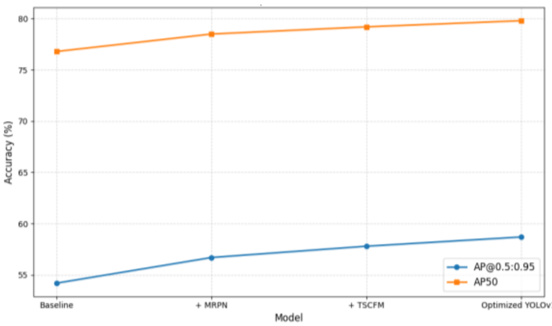

Figure 4. Performance Comparison

The ablation study results in Table 2 further confirm the contribution of each module. The MRPN module improves AP@0.5:0.95 by 2.5 percentage points (54.2% → 56.7%) and AP50 by 1.7 points (76.8% → 78.5%). The TSCFM module provides even greater improvements, increasing AP@0.5:0.95 by 3.6 points (54.2% → 57.8%) and AP50 by 2.4 points (76.8% → 79.2%). When both modules are applied, the overall AP@0.5:0.95 reaches 58.7%, outperforming the use of MRPN or TSCFM alone by 1.0 and 0.9 points, respectively. These results indicate that MRPN and TSCFM complement each other effectively: MRPN enhances multi-resolution feature representation, while TSCFM improves temporal–spatial consistency, jointly boosting keypoint detection accuracy without compromising real-time performance.

Table 2. Ablation study

Method	AP@0.5:0.95	AP50
Baseline	54.2%	76.8%
+ MRPN	56.7%	78.5%
+ TSCFM	57.8%	79.2%
Optimized YOLOv11	58.7%	79.8%

The effect diagrams of human keypoint detection are shown in Figures 5. The optimized YOLOv11 benefits from the combination of MRPN for robust multi-scale feature modeling and TSCFM for enhanced temporal–spatial context integration. This synergy significantly improves performance under rapid pose transitions and complex motion scenarios. While maintaining 48 FPS real-time inference speed, the model achieves 4.5-point and 3.0-point improvements in AP@0.5:0.95 and AP50, respectively. These results demonstrate that the proposed design effectively balances accuracy and efficiency, providing a high-performance solution for dynamic human pose estimation tasks.

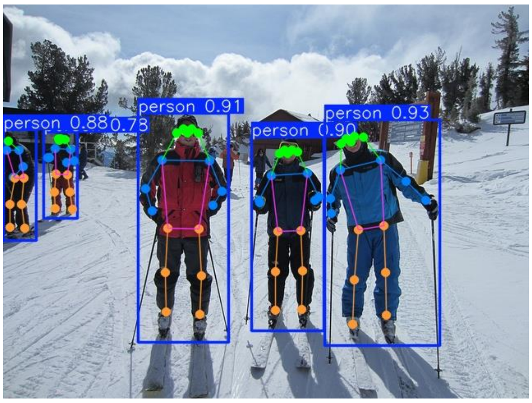

Figure 5. Detection Figure

5. Conclusion

In this work, we proposed an enhanced human keypoint detection framework based on YOLOv11, addressing two long-standing challenges in pose estimation: inaccurate multi-scale keypoint localization and temporal instability under dynamic motion. To overcome these limitations, we introduced two novel modules—the Multi-Resolution Parallel Network (MRPN) and the Temporal Spatial Context Fusion Module (TSCFM) — which respectively improve multi-scale feature representation and temporal–spatial consistency.

The MRPN module maintains parallel multi-resolution feature streams and enables effective cross-scale interaction, substantially enhancing the detection performance of both small and large keypoints. By preserving high-resolution representations throughout the network and introducing lightweight adaptive fusion connections, MRPN strengthens spatial feature completeness while maintaining computational efficiency.

The TSCFM module further improves robustness by integrating temporal continuity and spatial contextual cues. Through the Dynamic Temporal Modeling Unit, the system mitigates keypoint drift across frames, while the Spatial Context Enhancement Layer refines local joint representations. Together, these components ensure consistent and accurate keypoint localization even under rapid motion, occlusion, and complex pose variations.

Extensive experiments on the COCO 2017 keypoint validation set demonstrate the effectiveness of the proposed approach. The optimized YOLOv11 achieves an AP@0.5:0.95 of 58.7%, outperforming the baseline by 4.5 percentage points while maintaining real-time performance at 48 FPS. The ablation studies confirm that MRPN and TSCFM contribute complementary improvements, jointly delivering superior accuracy without compromising speed.

Overall, this work provides a practical and efficient solution for real-time human keypoint detection in dynamic environments. The proposed modules are lightweight, generalizable, and can be seamlessly integrated into other one-stage detectors, making the approach suitable for applications such as human–computer interaction, motion capture, intelligent surveillance, and sports analytics. Future research will explore incorporating transformer-based global modeling and extending the framework to 3D pose estimation and multi-person scenarios.

References

[1] Zhang J, Chen Z, Tao D. Towards high performance human keypoint detection. International Journal of Computer Vision, 2021, 129(9): 2639-2662.

[2] Geng Z, Sun K, Xiao B, et al. Bottom-up human pose estimation via disentangled keypoint regression //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14676-14686.

[3] McNally W, Vats K, Wong A, et al. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation //European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 37-54.

[4] He L, Zhou Y, Liu L, et al. Research on object detection and recognition in remote sensing images based on YOLOv11. Scientific Reports, 2025, 15(1): 14032.

[5] Cheng T, Song L, Ge Y, et al. Yolo-world: Real-time open-vocabulary object detection //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16901-16911.

[6] Wang W, Zhang J, Niu L, et al. Parallel multi-resolution fusion network for image inpainting //Proceedings of the IEEE/CVF international conference on computer vision. 2021: 14559-14568.

[7] Sheng X, Li L, Liu D, et al. Spatial decomposition and temporal fusion based inter prediction for learned video compression. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7): 6460-6473.

[8] Herpers R, Michaelis M, Lichtenauer K H, et al. Edge and keypoint detection in facial regions //Proceedings of the Second International Conference on Automatic Face and Gesture Recognition. IEEE, 1996: 212-217.

[9] Li Y, Zhang S, Wang Z, et al. Tokenpose: Learning keypoint tokens for human pose estimation //Proceedings of the IEEE/CVF International conference on computer vision. 2021: 11313-11322.

[10] Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression //European conference on computer vision. Cham: Springer International Publishing, 2016: 717-732.

[11] Yu C, Xiao B, Gao C, et al. Lite-hrnet: A lightweight high-resolution network //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10440-10450.

[12] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. Machine learning and knowledge extraction, 2023, 5(4): 1680-1716.

[13] Chen Y, Dai X, Chen D, et al. Mobile-former: Bridging mobilenet and transformer //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5270-5279.

[14] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection //Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[15] Howard M W, Fotedar M S, Datey A V, et al. The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. Psychological review, 2005, 112(1): 7