electronics

MDPI

Article

Context-Aware Human Pose Estimation via Hierarchical Information Arbitration

Jiayuan Wang, Jie Lv *, Xiaoru Chen and Yong Yang

School of Mechatronic Engineering, Guangdong Polytechnic Normal University, Guangzhou 510665, China; wjy@stu.gpnu.edu.cn (J.W.); chenxiaoru@stu.gpnu.edu.cn (X.C.); yy2008@gpnu.edu.cn (Y.Y.)
* Correspondence: lvjie@gpnu.edu.cn

Abstract

Human pose estimation requires accurate localization of body keypoints under complex backgrounds, occlusion, and diverse human postures. Existing high-resolution pose-estimation networks preserve spatial details effectively, but their static information flow limits their adaptability to different image contexts. To address this limitation, this paper proposes a context-aware hierarchical information arbitration method that dynamically regulates feature interaction at both multi-resolution fusion and residual feature refinement levels. The proposed method achieves superior performance on COCO, reaching 77.0 average precision and improving the High-Resolution Network baseline by 3.6 percentage points, with only a minor increase in model parameters. These results demonstrate that adaptive information arbitration improves pose-estimation accuracy and robustness while maintaining computational efficiency.

Keywords: human pose estimation; high-resolution networks; cross-branch attention; dynamic residual recalibration; multi-scale fusion

1. Introduction

Human pose estimation is a crucial task in computer vision, aiming to accurately localize key points on the human body from images or videos. It has a wide range of practical applications in fields such as human–computer interaction [1], motion analysis [2], and medical rehabilitation [3]. The primary goal of human pose estimation is to extract detailed information about key body parts by identifying coordinates and categories of individual key points (e.g., knees, elbows, shoulders, wrists) and constructing connections among these points, ultimately forming a representation of the human skeleton. Among various architectural innovations, High-Resolution Networks (HRNet) [4] marked a significant breakthrough. HRNet and its successors have consistently set high benchmarks for key point accuracy by maintaining high-resolution feature representations throughout the network via parallel multi-resolution branches. Concurrently, single-stage methods like KAPAO [5] and YOLO-Pose [6] have emerged, prioritizing inference speed by unifying detection and pose estimation. However, these real-time approaches often struggle to preserve fine-grained spatial details compared to top-down architectures, especially in crowded scenes.

However, beneath the success of these powerful models lies a fundamental, yet often overlooked, limitation: their internal information flow is predominantly static and context-agnostic. Most existing approaches employ a simplistic weighted-concatenation strategy [7] for multi-scale feature fusion; these methods fail to fully exploit complementary

Check for updates

Received: 23 April 2026
Revised: 15 May 2026
Accepted: 17 May 2026
Published: 20 May 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) license.

information among different branches (e.g., high-resolution versus low-resolution features), which causes the network to weaken critical information in complex scenarios. Methods in [8] help maintain high-resolution representations; however, they typically lack effective cross-scale and cross-branch feature interaction mechanisms, further limiting the accuracy of key point detection. Finally, traditional residual modules [9] utilize fixed, static fusion methods. They cannot adaptively adjust according to the global context of input images, thereby restricting the flexibility and expressiveness of feature representations. This rigid design implies that the network processes a simple image of a standing person like a complex scene with severe occlusion and intricate poses. This inability to adapt to the input creates a critical information flow bottleneck, leading to suboptimal feature representations in challenging real-world scenarios (e.g., crowded scenes as seen in Pose2Seg [10]) where the importance of local details versus global context can vary dramatically.

To address the above limitations, this study is guided by the following scientific and research questions. First, can context-aware information arbitration alleviate the static and context-agnostic information flow limitation in high-resolution human pose-estimation networks? Second, can adaptive inter-branch feature arbitration improve multi-resolution feature fusion under occlusion and complex backgrounds? Third, can Dynamic Residual Recalibration based on global feature differences improve residual feature refinement while maintaining computational efficiency?

To answer these questions, this paper proposes a context-aware hierarchical information arbitration framework. For the first question, the overall framework is designed to enable the network to dynamically regulate its internal information flow according to the semantic content of the input image. For the second question, we introduce a Cross-Branch Attention module to arbitrate multi-resolution feature fusion by dynamically evaluating the global importance of each resolution branch. For the third question, we design a Dynamic Residual Recalibration module to adaptively modulate residual connections by modeling the global feature difference between the transformed path and the shortcut path. These questions are further examined through comparative experiments, ablation studies, and visualization analyses in the experimental section.

Based on these research questions and the corresponding methodological design, the primary contributions of this work are summarized as follows:

1. We identify the static and context-agnostic information flow in HRNet-style architectures as a key limitation for pose estimation under occlusion and complex backgrounds and formulate hierarchical information arbitration as a unified solution at both inter-branch fusion and intra-block residual modulation levels.
2. We propose a Cross-Branch Attention (CBA) module that performs competition-based branch weighting for multi-resolution fusion, and a Dynamic Residual Recalibration (DRR) module that uses the global feature difference between the transformed and shortcut paths to adaptively regulate residual updates.
3. Extensive experiments on COCO 2017, MPII and CrowdPose show that CARPose improves the HRNet-W32 baseline by 3.6 AP on COCO with only a negligible parameter increase, demonstrating that the proposed arbitration mechanism enhances both accuracy and efficiency.

2. Related Work

Human pose estimation has been studied from multiple perspectives, including network architecture design, multi-scale feature fusion, and adaptive feature modulation. This section briefly reviews these related directions to clarify the motivation of the proposed method.

2.1. Deep Learning Based Human Pose Estimation

Deep learning has propelled rapid advancements in human pose estimation in recent years. DeepPose [11] pioneered convolutional neural networks (CNN) for human pose estimation by directly mapping images to the coordinates of the human key point. Despite its early successes, this method cannot capture multi-scale information. Subsequently, the Hourglass network [12] introduced a symmetric encoder–decoder architecture that captures multi-scale contextual information through iterative up-sampling and down-sampling, thereby improving key point accuracy. However, the sequential design of the Hourglass network can lose high-resolution details during up-sampling, particularly in complex scenes where local features are challenging to recover. To address these limitations, the Cascaded Pyramid Network (CPN) [13] implements a multi-stage pyramid structure for progressively refining the detection of challenging key points. Although CPN yields significant improvements in key point detection, its robustness remains poor under occlusion scenarios due to error accumulation across multiple stages. Meanwhile, transformer-based models [14] have demonstrated strong performance in capturing long-range dependencies and global context. Representative methods such as ViTPose [15] and TokenPose [16] incorporate transformer architectures into human pose estimation from different perspectives. ViTPose relies on plain vision transformer backbones to obtain scalable global representation learning, while TokenPose introduces learnable keypoints tokens to model the relationships between anatomical keypoints and image features. However, these methods usually depend heavily on transformer attention for global dependency modeling and may involve relatively high parameter or computational cost, especially when large transformer backbones are used. Moreover, they do not directly address the static multi-resolution fusion problem in HRNet-style high-resolution architectures.

Recent trends have also explored end-to-end transformer frameworks and real-time pose-estimation solutions. ED-Pose [17] and Group Pose [18] formulate pose estimation as a set-prediction problem, simplifying the multi-person pose-estimation pipeline by using instance-level and keypoint-level queries. However, these approaches mainly redesign the overall detection-estimation paradigm and generally require heavier training resources. On the other hand, RTMPose [19] and DWPose [20] push the limits of real-time performance through efficient architecture design and distillation, but they do not explicitly perform hierarchical information arbitration between multi-resolution branches and residual paths. In contrast, our CARPose aims to improve HRNet-style High-Resolution Network by introducing lightweight context-aware information flow control.

2.2. Multi-Scale Fusion and Cross-Branch Information Interaction

In proposing High-Resolution Network (HRNet) [4], researchers introduced a new era in human pose estimation. This method maintains high-resolution feature representations throughout the process using parallel multi-resolution subnetworks and cross-scale feature fusion, yielding excellent key point performance. However, HRNet typically employs a static fusion strategy by summing features pixel-wise across scales. Although this approach is structurally simple and computationally efficient, it fails to distinguish the dynamic importance of different branches adaptively. This leads to a loss of critical information in scenarios with drastic pose changes or severe occlusions. Methods such as Adaptively Spatial Feature Fusion (ASFF) [21] and Weighted Bidirectional Feature Pyramid Network (BiFPN) [7] address this limitation by weighting multi-scale features; ASFF uses a one-way fusion mechanism, whereas BiFPN introduces additional connections to enable dynamic weighting, resulting in significantly higher computational overhead. Similar context-aware and adaptive feature-interaction strategies have also been explored in related visual recognition tasks, including modality-aware RGB-T salient object detection and CLIP-driven

fine-grained text-image person re-identification [22,23]. Recently, some researchers have focused on deeper cross-branch interaction and attention-based high-resolution modeling. HRFormer [24] integrates the windowed attention mechanism from Swin Transformer [25] into a high-resolution architecture to enhance local–global information exchange, while Deformable Attention Transformer (DAT) [26] employs deformable attention to focus on relevant spatial regions dynamically. These transformer-based or attention-enhanced models improve the representation capacity of visual features, especially in complex scenes where long-range dependency and adaptive spatial focus are important.

However, their design focus is different from that of the proposed CBA module. HRFormer and DAT mainly enhance feature representation through self-attention or deformable attention, whereas CBA explicitly targets the multi-resolution branch-fusion process in HRNet-style architectures. Specifically, CBA learns competition-based weights for different resolution branches before feature fusion, allowing the network to decide whether high-resolution spatial details or low-resolution semantic context should be emphasized under different input conditions. Therefore, CARPose provides a lightweight branch-level arbitration mechanism rather than introducing transformer-heavy feature modeling.

2.3. Dynamic Feature Modulation and Residual Fusion

Traditional residual modules fuse the main branch and the residual path using simple summation. This fixed fusion strategy is stable in most scenarios; however, under significant variations in global image context or complex backgrounds, fixed weights cannot adequately adapt to the dynamic requirements of different regional features, limiting the model's ability to capture challenging key points. SENet [27] and ECANet [28] attempt to recalibrate feature responses by introducing channel attention mechanisms. However, these methods typically employ Sigmoid functions to model channel importance independently, effectively treating each channel as an isolated entity. Consequently, they lack the mechanism to explicitly model the competitive relationship required for optimal feature arbitration. To overcome static limitations, more advanced dynamic mechanisms have been proposed. Dynamic Convolution [29] and CondConv [30] demonstrate that adapting model parameters (e.g., convolution kernels) based on input features significantly boosts representational power. Similarly, SKNet [31] introduces a selective kernel mechanism to adjust receptive fields adaptively. However, these methods typically operate at the convolution or kernel level, which increases optimization complexity. To address the limitations of fixed weights in residual paths, Dynamic Residual Networks (DRN) [32] employ gating mechanisms to modulate shortcuts. However, these mechanisms typically generate weights based on local spatial features, thereby neglecting the global context required to resolve complex occlusions.

Different from these approaches, our DRR module leverages the global feature difference between the transformed branch and the shortcut branch. This design is directly motivated by residual learning: since the transformed branch represents a residual update relative to the shortcut reference, their global difference provides an explicit descriptor of residual displacement. Compared with single-path descriptors or concatenated descriptors, this differential signal offers a compact and contrastive cue for estimating whether the residual update should be emphasized or suppressed, enabling finer-grained residual fusion without substantially increasing computational complexity.

3. Method

The high-resolution human pose-estimation network proposed in this paper comprises a stem layer, transition layers, multi-stage parallel branches, fusion layers, and the proposed CBA and DRR modules, as shown in Figure 1. The multi-stage parallel branches denote

the multi-resolution streams within Stage 2, Stage 3, and Stage 4. The transition layers are located between adjacent stages and are used to adjust the number of branches, spatial resolutions, and channel dimensions. The fusion layers refer to the cross-resolution aggregation operations within each high-resolution module. The global average pooling and feature-alignment operations are performed inside the CBA blocks before multi-resolution feature fusion.

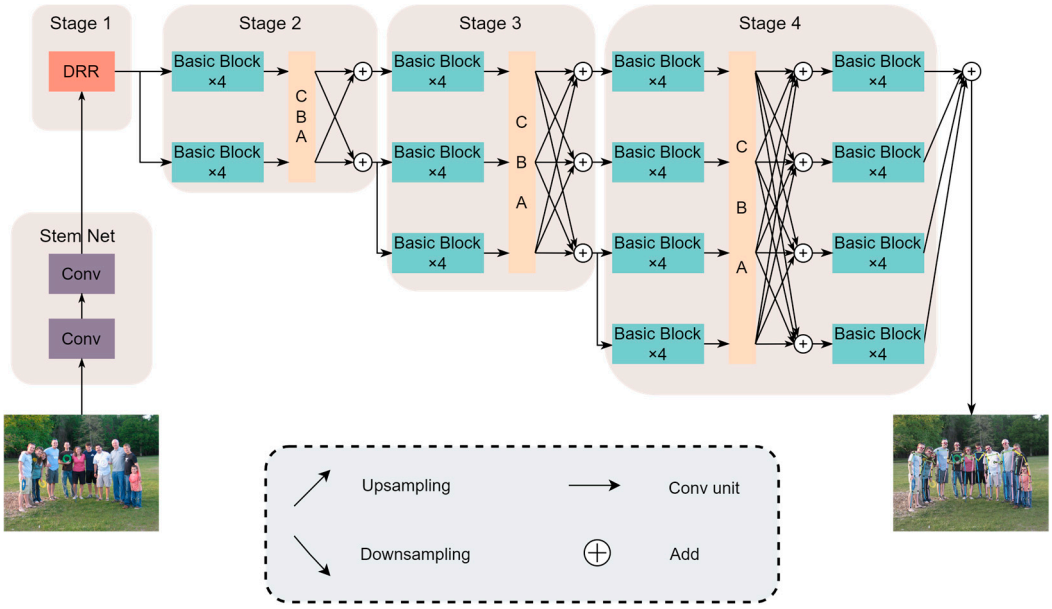


Figure 1. CARPose overall architecture.

We uniformly integrate a CBA module after each stage's parallel branch output to address the weakening of key information caused by fixed pixel-wise summation in traditional multi-scale feature fusion. The CBA module generates lightweight attention weights via global average pooling and feature alignment, enabling dynamic weighted fusion of different-resolution features to enhance each branch's key point discrimination capability adaptively. Furthermore, to overcome the fixed fusion limitation between the main branch and residual path in traditional Bottleneck residual modules, we introduce a DRR module, which leverages global context information to produce dynamic gating coefficients that allow residual fusion to adapt flexibly to input image characteristics. Our design improves accuracy and robustness in key points under complex occlusion and background interference while preserving high-resolution representation and low computational overhead.

3.1. Cross-Branch Attention Module

This study proposes a novel CBA module to fully leverage the complementary information from each branch in a multi-branch network. In the multi-resolution parallel architecture, each stage contains multiple parallel branches; each branch first extracts feature through a convolutional residual block, after which fusion layers align and aggregate features across different resolutions. In this paper, we insert the CBA module between the parallel branch outputs and the fusion layers for multi-resolution feature fusion, as shown in Figure 2. The designed CBA module comprises the following key steps:

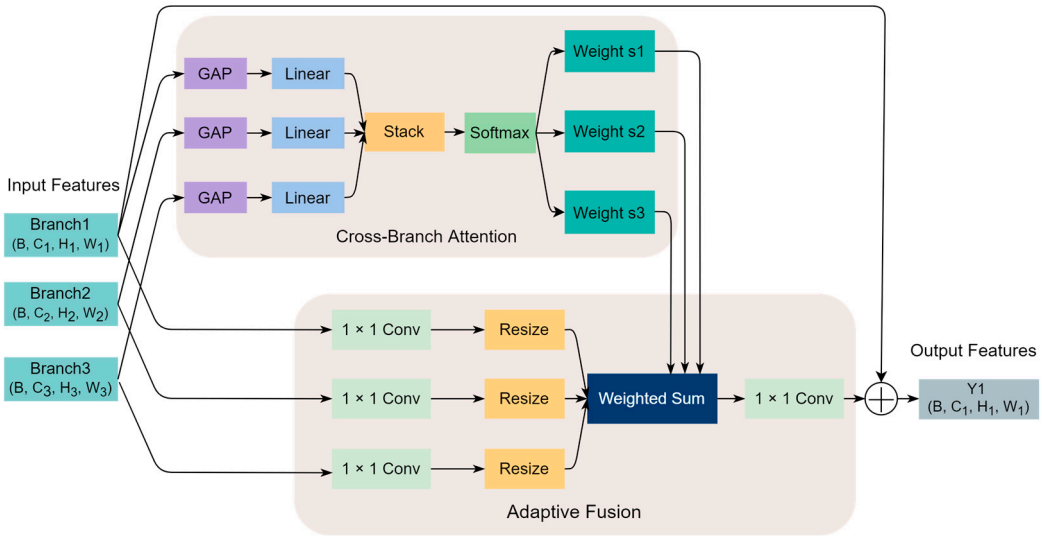


Figure 2. Cross-Branch Attention (CBA) module.

3.1.1. Global Description Vector Generation and Alignment

Given the feature map $X_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}, (i = 1, 2, \ldots, N)$, where B denotes the batch size, C_i is the number of channels, and H_i and W_i denote the spatial height and width, respectively, we first apply global average pooling to obtain a global feature vector

$$d_i = \mathrm{GAP}(X_i) \in \mathbb{R}^{B\times C_i} \tag{1}$$

Here, $\mathrm{GAP}(X_i)$ denotes global average pooling, which averages each feature channel over the spatial dimensions of the feature map to obtain a branch-level global descriptor.

To address the varying channel dimensions across branches, each di is then passed through a dedicated linear transformation $W_{d_i} \in \mathbb{R}^{C_i\times att}$ to map it to a common attention dimension *att* (set to 32 in this paper), yielding

$$\hat{d}_i = W_{d_i} d_i \in \mathbb{R}^{B\times att} \tag{2}$$

3.1.2. Weighting of Attention

We stack the aligned description vectors $\hat{d}_i = W_{d_i} d_i \in \mathbb{R}^{B\times att}$ from all branches into a tensor. We extract intermediate features for each branch via a shared fully connected layer W_f, apply a ReLU activation, and pass the result through a linear layer W_a to obtain raw scores. We then normalize these scores across branches using softmax [33] to compute the attention weights:

$$s_i = \mathrm{softmax}\Big(W_a \mathrm{ReLU}(W_f \hat{d}_i)\Big), i = 1, 2, \ldots, N, \tag{3}$$

where $s_i \in \mathbb{R}^{B\times 1}$ represents the contribution weight of branch i in the global feature fusion. Unlike standard channel attention modules (e.g., SE-Net) that use Sigmoid to calibrate features independently, the proposed CBA utilizes SoftMax to enforce a competitive mechanism. This explicitly arbitrates the information flow, ensuring that the network prioritizes the most semantically relevant resolution branch (e.g., high-resolution for details vs. low-resolution for context) rather than treating them in isolation.

3.1.3. Feature Alignment and Weighted Fusion

To ensure that each branch's feature map has the same number of channels, we first apply a 1×1 convolution W_{f_i} to the original feature $\hat{X}_i$ from branch i, mapping it to the common dimension att, yielding

$$\hat{X}_i = W_{f_i} X_i \in \mathbb{R}^{B \times att \times H_i \times W_i} \tag{4}$$

Next, for each target branch i, we resize the feature maps X_j from all branches to the spatial size (H_i, W_i) using bilinear interpolation. We weigh and sum up these resized features by their attention weights s_j and add the result to $\hat{X}_i$. Finally, we apply a 1×1 convolution W_{o_i} to reduce the fused feature map back to the original channel dimension C_i, yielding

$$Y_i = W_{o_i}\left(\hat{X}_i + \sum_{j=1}^{N} s_j \text{Resize}(X_j, H_i, W_i)\right) \in \mathbb{R}^{B \times C_i \times H_i \times W_i} \tag{5}$$

Overall, this design transcends simple feature recalibration. By introducing a competition-based arbitration mechanism via Softmax, the CBA module overcomes the static-weight limitation of traditional multi-resolution fusion. Unlike methods that process channels independently, CBA effectively routes the global context to the most critical resolution branch, enabling the network to adaptively balance between local details and global semantics under complex occlusion scenarios. Inserted between parallel branch outputs and fusion layers, it ensures that complementary information is dynamically enhanced prior to multi-scale summation.

3.2. Dynamic Residual Recalibration Module

To enable residual fusion to be flexibly adjusted according to the global context of the input image, rather than being performed through static summation, we propose a Dynamic Residual Recalibration (DRR) module, as shown in Figure 3. The design of DRR is motivated by the residual learning formulation. In a standard residual block, the shortcut branch provides a reference representation, while the transformed branch learns a residual update with respect to this reference. Therefore, residual modulation should depend not only on the transformed feature itself, but also on its deviation from the shortcut representation.

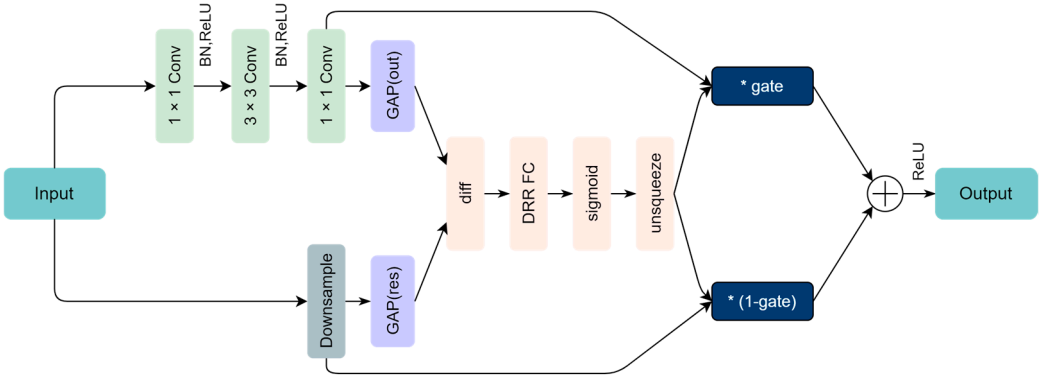


Figure 3. Dynamic Residual Recalibration (DRR) module. The symbol "*" denotes element-wise multiplication.

Let F denote the transformed feature produced by the convolutional branch and R denote the shortcut feature. After global average pooling, GAP(F) and GAP(R) summarize

the global semantic statistics of the transformed and reference representations, respectively. We then compute their differential descriptor as follows:

$$\Delta = \mathrm{GAP}(F) - \mathrm{GAP}(R) \tag{6}$$

The resulting vector Δ explicitly characterizes the global residual displacement introduced by the current block. Compared with using GAP(F) alone, Δ provides a reference-aware descriptor because it measures how far the transformed representation has moved away from the shortcut representation. GAP(F) alone only describes the transformed feature but does not indicate whether this transformation is necessary or excessive relative to the original residual input. Compared with using GAP(R) alone, Δ further captures the newly introduced residual update, whereas the shortcut descriptor only reflects the preserved identity information.

Compared with concatenation-based descriptors such as [GAP(F), GAP(R)], the subtraction formulation provides a compact reference-aware representation for the lightweight gating network. By directly encoding the global difference between the transformed branch and the shortcut branch, this descriptor captures the residual discrepancy without increasing the descriptor dimension. This design is consistent with the residual-learning formulation, in which the transformed branch can be regarded as an update relative to the shortcut reference. Therefore, Δ = GAP(F) − GAP(R) provides a simple and interpretable cue for estimating the necessity of residual modulation.

Compared with static residual summation, the proposed formulation enables input-dependent residual control, allowing the network to adjust the contribution of the transformed and shortcut branches according to the global feature discrepancy. This adaptive fusion mechanism helps suppress unreliable residual transformations caused by occlusion, background clutter, or ambiguous local patterns, while preserving useful pose-related refinements. Thus, Δ is introduced as a residual-learning-guided global discrepancy descriptor for Dynamic Residual Recalibration in complex pose-estimation scenarios.

3.2.1. Global Feature Variance Calculation

Given the main branch output $F \in \mathbb{R}^{B\times C\times H\times W}$ and the shortcut branch output $R \in \mathbb{R}^{B\times C\times H\times W}$ in the Bottleneck module, we first apply global average pooling to each tensor independently, yielding the global feature vectors

$$d_F = \mathrm{GAP}(F) \in \mathbb{R}^{B\times C}, d_R = \mathrm{GAP}(R) \in \mathbb{R}^{B\times C} \tag{7}$$

We then compute their difference:

$$\Delta = d_F - d_R \tag{8}$$

3.2.2. Dynamic Gating Factor Generation

We feed the difference vector Δ into a lightweight, fully connected network composed of two linear layers with a ReLU activation between them, followed by a Sigmoid function to produce the dynamic gating coefficients $g \in \mathbb{R}^{B\times C}$:

$$g = \sigma(W_2\mathrm{ReLU}(W_1\Delta)) \tag{9}$$

To ensure the fusion ratio is balanced, we initialize the network parameters with near-zero weights, making g approximately 0.5.

3.2.3. Dynamic Weighted Fusion

We then adaptively fuse the transformed branch output F and the shortcut output R using the dynamic gating coefficients to produce the module output:

$$Y = g \odot F + (1 - g) \odot R \tag{10}$$

where $\odot$ denotes element-wise multiplication. This mechanism enables the model to adjust information flow based on global feature variations in the input image, enhancing feature expressiveness and robustness. By incorporating the DRR module, we overcome the static fusion limitation of the traditional Bottleneck, enabling adaptive emphasis on critical information in complex scenes and significantly boosting pose-estimation accuracy.

3.3. Loss Function

To train the proposed CARPose network effectively, we employ a joint loss function that combines the Weighted Mean Squared Error (MSE) with an Online Hard Keypoint Mining (OHKM) strategy. This ensures that the model focuses on efficiently learning both visible easy key points and occluded hard key points.

3.3.1. Weighted Heatmap MSE Loss

Let $H_k \in \mathbb{R}^{W \times H}$ and $\hat{H}_k \in \mathbb{R}^{W \times H}$ denote the predicted heatmap and the ground–truth heatmap for the k-th key point, respectively, where K is the total number of key points (e.g., K = 17 for the COCO dataset). The ground–truth heatmaps are generated by applying a 2D Gaussian kernel centered on the ground–truth joint coordinates. The basic loss for the k-th key point is defined as:

$$L_k = \frac{1}{2} \sum_i \sum_j \parallel H_k(i,j) - \hat{H}_k(i,j) \parallel_2^2 \cdot V_k \tag{11}$$

where (i, j) represents the pixel location, and V_k is a visibility weight indicator. Specifically, V_k = 1 if the k-th key point is annotated (regardless of whether it is visible or occluded), and V_k = 0 if it is not annotated.

3.3.2. Online Hard Keypoint Mining (OHKM)

To address the challenge of "static information flow" where the network may neglect hard samples (e.g., heavily occluded joints) in favor of easier ones, we integrate the OHKM strategy. Instead of summing up the losses of all K key points directly, we dynamically select the top M key points with the highest losses during backpropagation.

Let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be the set of losses calculated for all key points in a single image. We sort these losses in descending order such that $L_{(1)} \geq L_{(2)} \geq \cdots \geq L_{(K)}$. The final total loss function L_{total} is formulated as:

$$L_{total} = \sum_{m=1}^{M} L_{(m)} \tag{12}$$

In our experiments, we set $M = 8$. Thus, for each training image, only the eight keypoints with the largest heatmap losses among the 17 COCO keypoints are included in the final loss used for backpropagation at each iteration. This mechanism forces the CARPose framework, particularly the CBA and DRR modules, to arbitrate information flow towards resolving these difficult pose ambiguities.

4. Experimental Testing and Analysis

In this section, we describe the experimental setup and training results, present ablation studies to validate our model's reliability, and provide visual analyses and comparisons to highlight the advantages of our Cross-Branch Attention mechanism and Dynamic Residual Recalibration module.

4.1. Experimental Environment

Experiments were conducted on Ubuntu 22.04 using a single NVIDIA GeForce RTX 4070 Ti Super GPU. We leveraged PyTorch 2.4.1 for deep learning and Python 3.8 for programming. COCO 2017 images were resized to 256×192 pixels during data preprocessing and augmented with random rotations (-45° to 45°), scaling factors from 0.65 to 1.35, horizontal flips, and half-body augmentation to improve robustness. For training, we leveraged the Adam optimizer with an initial learning rate of 1×10^{-3}, which decayed to 1×10^{-4} at epoch 170 and to 1×10^{-5} at epoch 200, training for a total of 210 epochs. During training on COCO 2017, the OHKM strategy described was applied with $M = 8$. Specifically, for each training image, losses were first computed for all 17 keypoints, and the 8 keypoints with the largest losses were selected to form the final loss for backpropagation, while the remaining keypoints were excluded from the gradient update in that iteration.

4.2. Dataset

We leverage three standard benchmarks: COCO 2017 [34] MPII [35] and CrowdPose [36] to evaluate our model's performance.

COCO 2017. The COCO 2017 dataset is a large-scale corpus of over 200,000 images and over 250,000 human instances annotated with 17 key points. It is split into training, validation (val 2017), and test-dev 2017 sets with approximately 57,000, 5000, and 20,000 images, respectively. Its scale, scene diversity, and challenging occlusions make it an excellent benchmark for assessing generalization and robustness. All experiments are trained exclusively on the training split and evaluated on val 2017 and test-dev 2017.

MPII. The MPII dataset from the Max Planck Institute for Informatics is a benchmark for single-person human pose estimation. It comprises over 25,000 images with over 40,000 human instances labeled with 16 key points. Featuring rich backgrounds, varied poses, and occlusions, MPII rigorously tests a model's ability to capture fine-grained pose details. By evaluating MPII, we further demonstrate the adaptability and robustness of our approach across different scenarios.

CrowdPose. To further evaluate the robustness of CARPose in crowded and occluded scenarios, we additionally conduct experiments on the CrowdPose benchmark. CrowdPose is designed for crowded human pose estimation and contains a large number of images with close human interactions, overlapping body regions, and inter-person occlusions. Compared with general-purpose pose-estimation datasets, CrowdPose provides a more focused evaluation setting for testing whether a model can maintain reliable keypoint localization under crowded-scene interference.

4.3. Evaluation Metric

The COCO 2017 dataset employs Object Keypoint Similarity (OKS) and Average Precision (AP) as primary evaluation metrics for key point detection, quantifying accuracy and robustness. OKS measures the normalized agreement between predicted and ground–

truth key points by applying a Gaussian penalty to the Euclidean distance, weighted by object scale and per-keypoint variance:

$$\mathrm{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{13}$$

where d_i is the Euclidean distance between predicted and ground–truth key point i, v_i the visibility flag, s the object's scale, k_i a per-keypoint constant controlling falloff, and $\delta(\mathrm{vi} > 0)$ an indicator equal to 1 if visible and zero otherwise.

AP is computed over multiple OKS thresholds (typically 0.50 to 0.95 in increments of 0.05) as the area under the precision–recall curve.

$$\mathrm{AP} = \frac{1}{P}\sum_{p=1}^{P} \delta\left(\mathrm{OKS}_p > T_p\right) \tag{14}$$

where P is the number of predictions, T_p the threshold for prediction p, and δ the indicator function. We report overall AP, AP^{50} and AP^{75}. Additionally, we include AP^{M} and AP^{L} to evaluate performance on medium and large objects, respectively. Average Recall (AR) measures recall averaged across the same set of OKS thresholds. These metrics comprehensively assess a pose-estimation model's performance across diverse scenarios and guide further optimization.

For the CrowdPose dataset, we follow the standard OKS-based evaluation protocol and report AP, AP50, and AP75 to evaluate the overall keypoints localization accuracy under crowded-scene conditions.

In MPII benchmarks, a commonly used evaluation metric is the Percentage of Correct Keypoints (PCK), together with its head-size-normalized variant, PCKh. Here, "head-size-normalized" means that the Euclidean distance between a predicted keypoint and its ground–truth position is divided by the head size of the corresponding person. Therefore, PCKh evaluates whether a predicted keypoint is sufficiently close to the ground–truth relative to the person's head scale, rather than according to an absolute pixel distance. This normalization reduces the influence of person-scale variation and enables fairer comparison across individuals of different sizes. For each key point, we compute the Euclidean distance between the predicted position pi and the ground–truth position g_i, then normalize this error by a reference scale h (e.g., head size or another body-part length):

$$e_i = \frac{\| p_i - g_i \|_2}{h} \tag{15}$$

Given a threshold α (set to 0.5 in this paper), meaning a prediction is correct if $e_i \leq \alpha$, we define the correctness indicator:

$$\mathrm{correct}_i = \begin{cases} 1 & \text{if } e_i \leq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

PCK (or PCKh) is then the fraction of key points correctly predicted:

$$\mathrm{PCK} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Correct}_i \tag{17}$$

where N is the number of key points in the sample, PCK metrics intuitively reflect an algorithm's key detection accuracy at various scales and serve as essential criteria for evaluating human pose-estimation methods.

4.4. Analysis and Discussion of Comparison Experiments

In experiments on the COCO 2017 dataset, we employ HRNet-W32 [4] as the backbone to evaluate our method. As shown in Table 1, our approach achieves an overall AP of 77.0, representing a 3.6 percentage-point improvement over the baseline. It also yields gains of 4.1, 3.9, 1.8, 1.2, and 0.7 percentage points in AP^{50}, AP^{75}, AP^{M}, AP^{L} and AR, respectively. Compared with recent transformer-based and high-resolution attention-based methods, CARPose shows a strong accuracy-efficiency balance. ViTPose-B [15], TokenPose-L [16], and HRFormer-B [24] achieve 75.8, 75.8, and 75.6 AP, respectively, with higher parameter or computational costs than CARPose. In contrast, CARPose achieves 77.0 AP with only 28.9 M parameters and 7.12 GFLOPs. It also outperforms RTMPose-l [19] in AP and achieves comparable accuracy to HRNeXt [37] under a smaller input size. These results indicate that CARPose effectively enhances HRNet-style pose estimation through lightweight CBA and DRR modules without relying on transformer-heavy architectures, as illustrated in Figure 4.

Table 1. Comparison of results on the COCO2017 dataset.

Method	Backbone	Input Size	Params(M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stageHourglass [12]	Hourglass	256 × 192	25.1	14.3	66.9	-	-	-	-	-
CPN [13]	ResNet-50	256 × 192	27.0	6.20	68.6	-	-	-	-	-
SimpleBaseline [38]	ResNet-50	256 × 192	34.0	8.90	70.4	88.6	78.3	67.1	77.2	76.3
DARK [39]	HRNet-W32	128 × 96	28.5	1.80	70.7	88.9	78.4	67.9	76.6	76.7
PRTR [40]	HRNet-W32	384 × 288	57.2	21.6	73.1	89.4	79.8	68.8	80.4	79.8
FastNet [41]	Lite-HRNet	256 × 192	19.7	5.8	73.3	89.4	79.7	69.5	80.5	78.7
HRNet [4]	HRNet-W32	256 × 192	28.5	7.10	73.4	89.5	80.7	70.2	80.1	78.9
UDP [42]	HRNet-W32	256 × 192	28.7	7.10	75.2	92.4	82.9	72.0	80.8	80.4
ViTPose [15]	ViTPose-B	256 × 192	86.0	17.1	75.8	90.7	83.2	68.7	78.4	81.1
HRFormer-B [24]	HRFormer-B	256 × 192	43.2	12.2	75.6	90.8	82.8	71.7	82.6	80.8
TokenPose-L [16]	HRNet-W48	256 × 192	27.5	11.0	75.8	90.3	82.5	72.3	82.7	80.9
ConvPose-BL [43]	HRNet-W32	256 × 192	29.0	10.7	76.0	93.5	83.6	73.1	80.3	78.9
FDAPose [44]	HRNet-W48	256 × 192	27.2	10.4	76.5	93.6	83.7	73.5	81.1	79.2
RTMPose-l [19]	CSPNeXt-l	256 × 192	-	4.16	76.6	-	-	-	-	-
HRNeXt [37]	HRNeXt-B	384 × 288	31.7	10.8	77.0	91.1	83.5	73.2	84.1	82.0
Ours	HRNet-W32	256 × 192	28.9	7.12	77.0	93.6	84.6	74.0	81.3	79.6

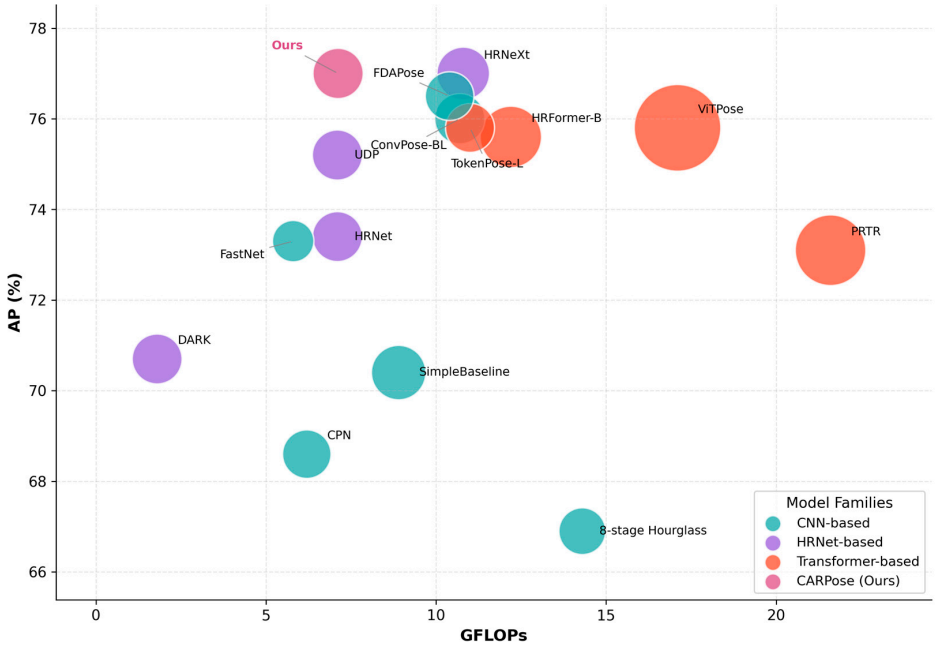


Figure 4. CARPose and other models on the COCO2017 dataset in accuracy and computational complexity; bubble size indicates parameter count.

In terms of computational complexity, CARPose introduces only a negligible overhead compared with HRNet-W32, increasing the number of parameters from 28.5 M to 28.9 M and GFLOPs from 7.10 to 7.12. Since actual inference time is highly dependent on the hardware platform, batch size, detector implementation, preprocessing, and post-processing, we follow the common practice of reporting Params and GFLOPs as hardware-independent indicators of model efficiency. Device-specific latency evaluation will be further investigated in future deployment-oriented work.

Table 2 compares our approach with SimpleBaseline [38] and the original HRNet-W32 on the MPII validation set. Although SimpleBaseline has 68.6 M parameters, it achieves only 91.5% PCKh@0.5. HRNet-W32 reduces the parameter count to 28.5 M and raises PCKh@0.5 to 92.3%. Our method leverages 28.9 M parameters, just 0.4 M more than HRNet-W32, with comparable GFLOPs (9.6 vs. 9.5) and further boosts PCKh@0.5 to 93.2%. These results demonstrate that our CBA and DRR modules enhance multi-scale feature fusion and dynamic residual modulation without significantly increasing model complexity, achieving higher key point localization accuracy on MPII.

Table 2. Comparison of results on the MPII dataset.

Method	Params (M)	GFLOPs	PCKh
DeeperCut [45]	42.6	41.2	88.5
Hourglass [12]	25.1	19.1	90.9
SimpleBaseline [38]	68.6	20.9	91.5
PyraNet [46]	28.1	21.3	92.0
DLCM [47]	15.5	15.6	92.3
HRNet-W32 [4]	28.5	9.5	92.3
Ours	28.9	9.6	93.2

To further validate the effectiveness of CARPose under crowded-scene conditions, we additionally evaluate the proposed model on the CrowdPose dataset and compare it with representative existing methods. As shown in Table 3, CARPose achieves 81.9 AP, 94.7 AP50, and 87.0 AP75. Compared with traditional top-down methods such as Mask R-CNN and AlphaPose, CARPose obtains a clear improvement in overall AP. It also outperforms refinement-based SPPE and recent bottom-up methods such as HigherHRNet-W48+, DEKR-HRNet-W48, and BAPose-W32. These results demonstrate that the proposed CBA and DRR modules enhance HRNet-style high-resolution pose estimation under crowded-scene interference and further support the robustness of context-aware information arbitration in challenging multi-person scenarios.

Table 3. Comparison of results on the CrowdPose dataset.

Method	Backbone	Input Size	AP	AP^{50}	AP^{75}
Mask R-CNN [48]	ResNet-FPN	-	57.2	83.5	60.3
AlphaPose [49]	ResNet	-	61.0	81.3	66.0
SPPE [36]	-	-	66.0	84.2	71.5
HigherHRNet [50]	HRNet-W48	640	67.6	87.4	72.6
DEKR [39]	HRNet-W48	640	68.0	85.5	73.4
BAPose [51]	HRNet-W32	512	72.2	89.6	78.0
Ours	HRNet-W32	256 × 192	81.9	94.7	87.0

4.5. Ablation Study

Hyperparameters in CBA. To determine the optimal combination of att_channels count and reduction ratio in the CBA module, we performed nine ablation experiments on the COCO val 2017 dataset. The results are summarized in Table 4, with optimal settings

indicated by an asterisk. It is noteworthy that when attention channels and reduction ratio were set to 128 and 16, respectively, the model lost its discriminative power by the 88th epoch, and metrics such as AP and AP50 fell to zero. We attribute this collapse to a tiny hidden-layer dimension ($128 \div 16 = 8$), which impedes the network's ability to learn distinct branch attention weights; consequently, weight distributions become uniform or saturated, causing feature fusion to fail. In contrast, configuring the module with 32 attention channels and a reduction ratio of 4 yields a hidden dimension of 8. However, this $4\times$ compression is moderate, preserving more feature pathways and stable gradient flow. Conversely, the $16\times$ compression ($128\rightarrow8$) creates an extreme information bottleneck that prevents the attention module from differentiating branch-specific features and "squeezes" gradient signals, thereby impairing downstream updates. Therefore, selecting appropriate attention-channel counts and reduction ratios is essential for balancing model compactness, accuracy, and stability.

Table 4. Ablation study of attention-channel counts and reduction ratios in the CBA module.

Att_Channels	Reduction	Params (M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
128	4	29.64	7.45	76.7	93.6	84.6	73.9	81.0	79.4
128	8	29.62	7.45	76.7	93.6	84.7	73.8	81.0	79.4
128	16	29.61	7.45	-	-	-	-	-	-
64	4	29.14	7.29	76.5	93.6	83.7	73.7	80.9	79.3
64	8	29.14	7.29	76.5	93.6	83.7	74.1	81.0	79.4
64	16	29.14	7.29	76.7	93.5	83.7	74.0	81.0	79.5
32	4	28.90	7.20	76.7	93.5	83.7	74.1	81.1	79.5
32	8	28.90	7.20	77.0	93.6	84.6	74.0	81.3	79.6
32	16	28.90	7.20	76.7	93.6	84.6	74.0	81.0	79.5

Hyperparameters in DRR. To evaluate the structural sensitivity of the lightweight DRR-gating network, we varied the hidden layer dimension of the gating MLP on the COCO val 2017 dataset. In this experiment, the residual-discrepancy descriptor was fixed as $\Delta = \mathrm{GAP}(F) - \mathrm{GAP}(R)$, and only the hidden layer dimension was changed. Table 5 summarizes the performance of the DRR module under different hidden layer dimensions.

Table 5. Ablation study of different hidden layer dimensions in the DRR module.

Hidden Layer Dimension	Params (M)	GFLOPs	AP
2	29.03	7.20	76.7
4	28.90	7.20	77.0
8	28.84	7.20	76.5

Validating the effectiveness of the CBA and DRR. To further validate the effectiveness of our proposed modules, we performed an ablation study on the COCO val 2017 dataset. We added the CBA and DRR modules individually and in combination to evaluate their impact on model parameters, computational overhead, and average precision. Table 6 summarizes the detailed ablation results. Our approach achieves significant improvements compared to the baseline, confirming each module's contribution and highlighting their complementary roles in multi-scale feature interaction and residual-path modulation.

Table 6. Evaluation results of the effect of each component on the performance of CARPose obtained through an ablation study.

Method	DRR	CBA	Params (M)	GFLOPs	AP
CARPose	√	×	28.7	7.12	76.4
CARPose	×	√	28.8	7.20	76.8
CARPose	√	√	28.9	7.20	77.0

4.6. Internal Mechanism Analysis and Visualization

To intuitively understand how CARPose arbitrates information flow, we visualize the internal behaviors of the CBA and DRR modules on the COCO validation set.

Analysis of CBA Weights. The CBA module is designed to dynamically assign importance weights to different resolution branches. To verify this, we statistically analyzed the attention weights (s_i) generated for the high-resolution branch. To define sample difficulty, we ranked person instances in the COCO validation set by the visible-keypoint ratio r = Nvis/Nann, where Nvis and Nann denote the numbers of visible and annotated keypoints, respectively; the top 30% were regarded as easy samples, and the bottom 30% were regarded as hard samples. As illustrated in Figure 5, the median high-resolution branch weight s_1 is approximately 0.33 for easy samples and 0.68 for hard samples, and the corresponding interquartile ranges are approximately 0.25–0.40 and 0.61–0.74, respectively. This quantitative difference indicates that the CBA module assigns greater importance to high-resolution features when processing ambiguous or occluded samples. This distribution shift indicates that the CBA module successfully learns to prioritize detailed spatial information when the context is ambiguous.

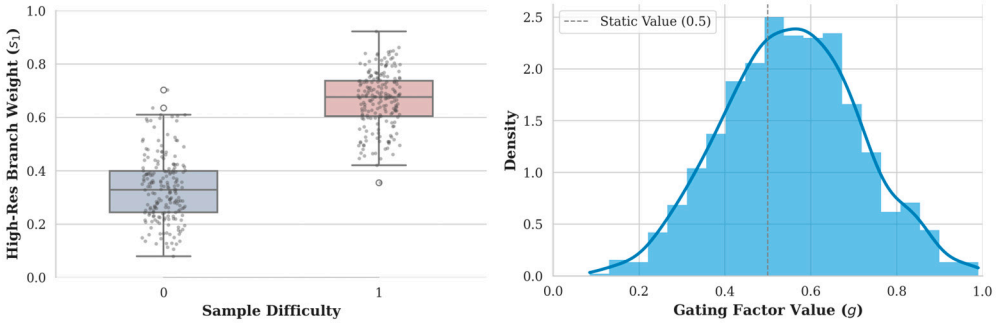


Figure 5. Visualization of internal mechanism analysis.

Analysis of the DRR-Gating Factor. The DRR module modulates the residual connection via a gating factor g. We visualized the density distribution of g values in Figure 5. The results show that g covers a dynamic range rather than remaining static at 0.5. Specifically, the gating values adaptively shift based on the magnitude of the feature difference Δ. This observation further supports the role of the feature difference signal in guiding active information regulation.

4.7. Qualitative Analysis

Qualitative predictions of our approach on the COCO 2017 dataset are presented in Figure 6, demonstrating adaptability and accuracy across diverse scenes. We evaluate four representative scenarios: (a) single person, (b) multiple people, (c) complex actions, and (d) partial occlusion. In the single person case (a), our model accurately localizes key points despite background complexity, whether cluttered or straightforward, and diverse poses, generating clear skeletal connections for standing, walking, or outdoor activities. The multiple-people scenario (b) validates stability in crowded settings: even

with close interactions, the model delineates individual skeletons without overlap or keypoint confusion. For scenario (c), our model tracks challenging movements such as jumps, flips, and irregular postures, producing skeletal structures that accurately reflect anatomy under extreme pose variations and rapid motion. Finally, for the partial-occlusion scenario (d), our model recovers visible joints when limbs or the torso are obscured by objects or the background, preserving overall skeletal integrity and maintaining high accuracy under realistic shooting conditions. Together, these examples demonstrate our model's resilience to interference and its adaptability to varying group sizes, occlusion levels, and movement scales, laying a solid foundation for further research and real-world application of human pose estimation.

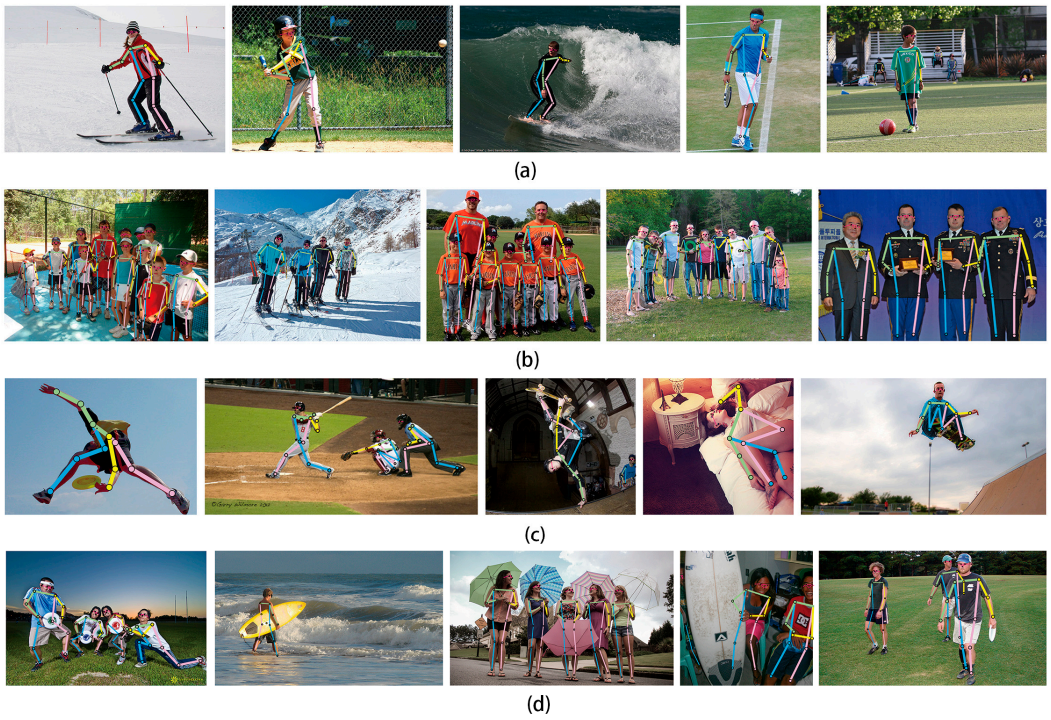


Figure 6. Visualizations on the COCO 2017 dataset: (**a**) single-person scene, (**b**) multi-person scene, (**c**) complex action scene, (**d**) partially occluded scene.

A detailed qualitative comparison between HRNet and our CARPose model on the COCO 2017 dataset is presented in Figure 7 CARPose consistently produces more accurate and coherent key point predictions than HRNet across challenging conditions, including overlapping subjects, occlusions, complex backgrounds, and unconventional poses, achieving finer joint localization and more reliable pose-structure recovery. These results underscore the benefits of incorporating Cross-Branch Attention and Dynamic Residual Recalibration modules in complex environments. This visual comparison intuitively highlights CARPose's improved prediction accuracy over HRNet.

To provide a more balanced qualitative analysis, Figure 8 presents representative failure cases of CARPose on the COCO 2017 dataset. These cases mainly involve dense multi-person overlap, body truncation, uncommon viewpoints, environmental occlusion, and interacting persons with motion ambiguity. Under such conditions, the visible body evidence may be incomplete or highly ambiguous, causing some keypoints to be missed, inaccurately localized, or incorrectly associated with nearby persons. These examples indicate that, although CARPose improves robustness in many complex scenes, pose estimation under extreme occlusion and ambiguous human interactions remains challenging.

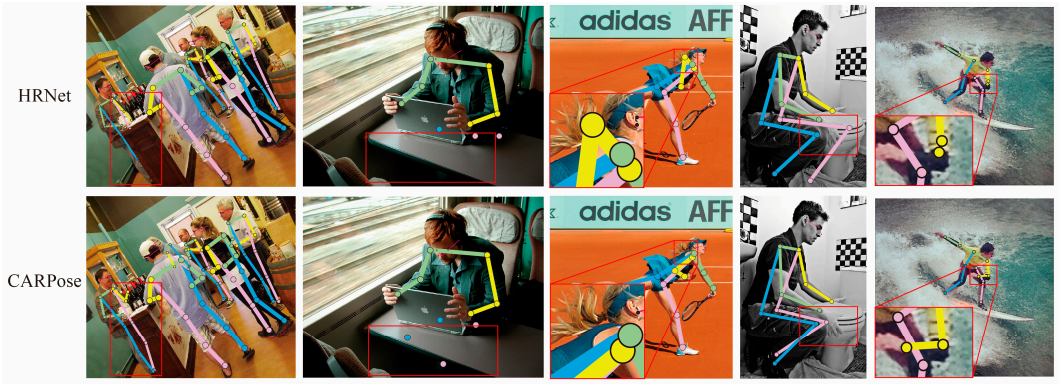


Figure 7. Comparison of CARPose and HRNet visualization predictions on the COCO 2017 dataset.



Figure 8. Representative failure cases of CARPose on the COCO 2017 dataset. (**a**) Dense multi-person overlap; (**b**) body truncation; (**c**) environmental occlusion; (**d**) interacting persons with motion ambiguity.

5. Conclusions and Future Works

In this paper, we addressed a fundamental limitation in high-resolution human pose-estimation networks: the static information flow bottleneck. We argued that the rigid, context-agnostic fusion strategies in existing models hinder their performance in complex, real-world scenarios. To resolve this, we introduced a novel context-aware hierarchical information arbitration framework, which empowers the network to intelligently and dynamically control its internal information flow. Our framework was instantiated in the CARPose architecture through two complementary modules: the Cross-Branch Attention module for macro-level arbitration of multi-scale features, and the Dynamic Residual Recalibration module for micro level modulation of residual connections.

Experimental results on COCO, MPII, and CrowdPose demonstrate that the proposed method effectively improves pose-estimation accuracy and robustness, particularly in challenging scenarios involving occlusion, complex backgrounds, and diverse postures. These results answer the research questions raised in the Introduction by showing that context-aware information arbitration can alleviate static information flow, adaptive inter-branch fusion can enhance multi-resolution feature interaction, and Dynamic Residual Recalibration can refine residual representations with only a minor increase in computational cost. Therefore, the proposed hierarchical arbitration strategy provides an effective and efficient design direction for high-resolution human pose estimation, with the added value of improving information flow in a lightweight and interpretable manner. Its implementation mainly required careful control of attention stability, hyperparameter settings, and computational cost.

Despite these advances, several avenues remain for future exploration. First, while our CBA module effectively fuses features, future work could explore incorporating more sophisticated contextual cues beyond global pooling to further enhance its arbitration capabilities in extremely cluttered scenes. Second, the DRR module could be extended to model local, spatial variations in the feature differential signal, potentially enabling spatially adaptive residual modulation. Finally, we plan to investigate model compression, quantization, and device-specific latency optimization to deploy our framework on

resource-constrained platforms, balancing the trade-off between accuracy and real-time performance for broader real-world applicability.

Author Contributions: J.W.: Writing—original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. J.L.: Writing—review and editing, Validation, Supervision, Resources, Methodology, Conceptualization. X.C.: Data curation. Y.Y.: Resources, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China Joint Fund Key Projects (No. U22A2062), Major Science and Technology R&D Special Project of Jiangxi Province (No. 20223AAE02008), Special Programs for Key Areas of Guangdong Universities (No. 2021ZDZX4039), Scientific Research Capacity Enhancement Program for Key Construction Disciplines in Guangdong Province (No. 2022ZDJS012), Guangdong Provincial Key Laboratory of Renewable Energy (No. E539kf0901) and Guangdong Polytechnic Normal University (No. 2025SDKYA006).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in https://data.mendeley.com/preview/8tfdvgcwkb?a=1947834c-4208-4475-ac64-8bfc9183d08e, accessed on 16 May 2026, also available on request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average Precision
AP^{50}	Average Precision at an Object Keypoint Similarity threshold of 0.50
AP^{75}	Average Precision at an Object Keypoint Similarity threshold of 0.75
AP^{M}	Average Precision for medium-scale persons
AP^{L}	Average Precision for large-scale persons
AR	Average Recall
CARPose	Cross-Branch Attention and Residual Recalibration Pose
CBA	Cross-Branch Attention
COCO	Common Objects in Context
CNN	Convolutional Neural Network
DRR	Dynamic Residual Recalibration
GFLOPs	Giga Floating-Point Operations
HRNet	High-Resolution Network
MPII	Max Planck Institute for Informatics
MSE	Mean Squared Error
OHKM	Online Hard Keypoint Mining
OKS	Object Keypoint Similarity
PCK	Percentage of Correct Keypoints
PCKh	Percentage of Correct Keypoints normalized by head size

References

1. Liu, H.; Liu, T.; Zhang, Z.; Sangaiah, A.K.; Yang, B.; Li, Y. ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7107–7117. [CrossRef]
2. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2959–2968.
3. Palermo, M.; Moccia, S.; Migliorelli, L.; Frontoni, E.; Santos, C.P. Real-time human pose estimation on a smart walker using convolutional neural networks. *Expert Syst. Appl.* **2021**, *184*, 115498. [CrossRef]
4. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696.

5. McNally, W.; Vats, K.; Wong, A.; McPhee, J. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 37–54.
6. 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 2636–2645.
7. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
8. Li, Y.; Hao, M.; Di, Z.; Gundavarapu, N.B.; Wang, X. Test-time personalization with a transformer for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2583–2597.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
10. Zhang, S.-H.; Li, R.; Dong, X.; Rosin, P.; Cai, Z.; Han, X.; Yang, D.; Huang, H.; Hu, S.-M. Pose2Seg: Detection free human instance segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 889–898.
11. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
12. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
13. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; N.Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
15. Tao, D.; Xu, Y.; Zhang, J.; Zhang, Q. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In Proceedings of the Advances in Neural Information Processing Systems 35, New Orleans, LA, USA, 28 November–9 December 2022; pp. 38571–38584.
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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11293–11302.
17. Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; Zhang, L. Explicit box detection unifies end-to-end multi-person pose estimation. *arXiv* **2023**, arXiv:2302.01593.
18. Liu, H.; Chen, Q.; Tan, Z.; Liu, J.-J.; Wang, J.; Su, X.; Li, X.; Yao, K.; Han, J.; Ding, E.; et al. Group pose: A simple baseline for end-to-end multi-person pose estimation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 14983–14992.
19. 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.
20. Yang, Z.; Zeng, A.; Yuan, C.; Li, Y. Effective whole-body pose estimation with two-stages distillation. *arXiv* **2023**, arXiv:2307.15880.
21. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516. [CrossRef]
22. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1958–1974. [CrossRef] [PubMed]
23. Yan, S.; Dong, N.; Zhang, L.; Tang, J. CLIP-Driven Fine-Grained Text-Image Person Re-Identification. *IEEE Trans. Image Process.* **2023**, *32*, 6032–6046. [CrossRef] [PubMed]
24. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. HRFormer: High-Resolution Transformer for dense prediction. *arXiv* **2021**, arXiv:2110.09408. [CrossRef]
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
26. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision Transformer with Deformable Attention. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4794–4803.
28. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.

29. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11030–11039.
30. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. *arXiv* **2019**, arXiv:1904.04971.
31. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
32. Figurnov, M.; Collins, M.D.; Zhu, Y.; Zhang, L.; Huang, J.; Vetrov, D.; Salakhutdinov, R. Spatially adaptive computation time for residual networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1039–1048.
33. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
34. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
35. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
36. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; Lu, C. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10863–10872.
37. Li, Q.; Zhang, Z.; Zhang, F.; Xiao, F. HRNeXt: High-resolution context network for crowd pose estimation. *IEEE Trans. Multimed.* **2023**, *25*, 1521–1528. [CrossRef]
38. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 472–487.
39. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7093–7102.
40. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose Recognition with Cascade Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1944–1953.
41. Luo, Y.; Ou, Z.; Wan, T.; Guo, J.-M. FastNet: Fast high-resolution network for human pose estimation. *Image Vis. Comput.* **2022**, *119*, 104390. [CrossRef]
42. Huang, J.; Zhu, Z.; Guo, F.; Huang, G. The devil is in the details: Delving into unbiased data processing for human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5700–5709.
43. Niu, Y.; Wang, A.; Wang, X.; Wu, S. ConvPose: A modern pure ConvNet for human pose estimation. *Neurocomputing* **2023**, *544*, 126301. [CrossRef]
44. Zhou, S.; Duan, X.; Zhou, J. Human pose estimation based on frequency domain and attention module. *Neurocomputing* **2024**, *604*, 128318. [CrossRef]
45. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 34–50.
46. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1281–1290.
47. Tang, W.; Yu, P.; Wu, Y. Deeply learned compositional models for human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 190–206.
48. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
49. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
50. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5385–5394.
51. Artacho, B.; Savakis, A. Full-BAPose: Bottom up framework for full body pose estimation. *Sensors* **2023**, *23*, 3725. [CrossRef] [PubMed]

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.