electronics

MDPI

Article

Attentive Multi-Scale Features with Adaptive Context PoseResNet for Resource-Efficient Human Pose Estimation

Ali Zakir [1], Sartaj Ahmed Salman [1], Gibran Benitez-Garcia [1,*] and Hiroki Takahashi [1,2]

[1] Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofugaoka 1-5-1, Chofu 182-8585, Japan; a2240012@edu.cc.uec.ac.jp (A.Z.); s2140019@edu.cc.uec.ac.jp (S.A.S.); rocky@inf.uec.ac.jp (H.T.)

[2] AI Exploration/Meta-Networking Research Center, The University of Electro-Communications, Chofugaoka 1-5-1, Chofu 182-8585, Japan

* Correspondence: gibran@ieee.org

Abstract: Human Pose Estimation (HPE) remains challenging due to scale variation, occlusion, and high computational costs. Standard methods often struggle to capture detailed spatial information when keypoints are obscured, and they typically rely on computationally expensive deconvolution layers for upsampling, making them inefficient for real-time or resource-constrained scenarios. We propose AMFACPose (Attentive Multi-scale Features with Adaptive Context PoseResNet) to address these limitations. Specifically, our architecture incorporates Coordinate Convolution 2D (CoordConv2d) to retain explicit spatial context, alleviating the loss of coordinate information in conventional convolutions. To reduce computational overhead while maintaining accuracy, we utilize Depthwise Separable Convolutions (DSCs), separating spatial and pointwise operations. At the core of our approach is an Adaptive Feature Pyramid Network (AFPN), which replaces costly deconvolution-based upsampling by efficiently aggregating multi-scale features to handle diverse human poses and body sizes. We further introduce Dual-Gate Context Blocks (DGCBs) that refine global context to manage partial occlusions and cluttered backgrounds. The model integrates Squeeze-and-Excitation (SE) blocks and the Spatial–Channel Refinement Module (SCRM) to emphasize the most informative feature channels and spatial regions, which is particularly beneficial for occluded or overlapping keypoints. For precise keypoint localization, we replace dense heatmap predictions with coordinate classification using Multi-Layer Perceptron (MLP) heads. Experiments on the COCO and CrowdPose datasets demonstrate that AMFACPose surpasses the existing 2D HPE methods in both accuracy and computational efficiency. Moreover, our implementation on edge devices achieves real-time performance while preserving high accuracy, confirming the suitability of AMFACPose for resource-constrained pose estimation in both benchmark and real-world environments.

Keywords: human pose estimation; multi-scale features; edge computing; dual-gate context blocks; adaptive feature pyramid network

Academic Editor: Stefanos Kollias

Received: 1 April 2025
Revised: 15 May 2025
Accepted: 20 May 2025
Published: 22 May 2025

Citation: Zakir, A.; Salman, S.A.; Benitez-Garcia, G.; Takahashi, H. Attentive Multi-Scale Features with Adaptive Context PoseResNet for Resource-Efficient Human Pose Estimation. *Electronics* **2025**, *14*, 2107. https://doi.org/10.3390/electronics14112107

1. Introduction

Human Pose Estimation (HPE) is a critical subfield of Computer Vision (CV) focused on locating and connecting human body joints, i.e., keypoints, in images or video sequences. Accurate pose estimation enables machines to analyze human posture and motion, facilitating a wide range of applications, including video surveillance, human–computer interaction, medical rehabilitation, and autonomous driving [1]. Despite significant advancements

in deep learning (DL) and convolutional neural networks (CNNs) [2], HPE continues to face several core challenges. These include scale variation, where subjects can appear at vastly different resolutions, the occlusion of keypoints by objects or other individuals, and stringent computational requirements that frequently hinder real-time deployment [3,4].

Early approaches relied on hand-crafted features combined with probabilistic models, offering interpretability but limited accuracy in complex scenarios, such as occlusions, varied lighting conditions, and cluttered backgrounds [5]. The emergence of CNNs significantly improved feature extraction capabilities. For instance, the pioneering work of DeepPose directly regressed keypoint coordinates but struggled with stability in multi-modal distributions [4]. Subsequently, heatmap-based methods emerged, transforming coordinate regression into spatial heatmap prediction, notably improving robustness and accuracy. Approaches such as Stacked Hourglass Networks [6] and SimpleBaseline [7] refined heatmap-based architectures, although at the expense of increased computational overhead due to complex upsampling procedures.

To reduce these computational demands while maintaining accuracy, High-Resolution Networks (HRNets) preserve spatial resolution throughout the network, significantly improving precision, albeit with substantial computational complexity [8]. Recognizing this trade-off, transformer-based models such as HRFormer [9], TokenPose [10], and ViTPose [11] leveraged global self-attention mechanisms, enabling more accurate keypoint estimation by modeling global contextual relationships. However, these approaches often introduced even greater computational complexity and parameter demands, making practical deployment challenging.

To address these efficiency challenges, coordinate classification approaches have recently emerged, such as SimCC [12], reformulating keypoint localization as a discrete classification task, significantly reducing the quantization errors inherent in heatmap-based methods. Building on this paradigm, AECA-PRNetCC further enhanced performance by incorporating adaptive channel attention mechanisms, thereby achieving a balance between accuracy and computational efficiency [13].

Despite these improvements, two major issues persist. Firstly, most advanced methods face a fundamental accuracy–computation trade-off, as higher accuracy typically demands greater computational complexity, complicating real-world deployment. Secondly, existing approaches still suffer from precision limitations in localizing keypoints under occlusion or scale variations, partly due to ineffective local feature refinement and global contextual reasoning capabilities.

In response, we propose AMFACPose (Attentive Multi-scale Features with Adaptive Context PoseResNet), a novel framework designed to address these challenges in 2D HPE. Our model begins with a ResNet structure [14], which we modified by replacing the standard convolution layers with Coordinate Convolution 2D (CoordConv2d) [15] to preserve explicit spatial coordinates, as well as by removing the average pooling and fully connected layers. This design retains the model's feature-extraction capabilities while reducing computational overhead. We also replace the standard 7×7 convolution in the initial layer with a series of 3×3 CoordConv2d layers, each followed by Batch Normalization (BN) and Mish activation, thereby improving the model's ability to capture fine-grained features. Furthermore, throughout the four stages of ResBlocks, we employ Depthwise Separable Convolutions (DSCs) [16] to further reduce computational costs without compromising accuracy, separating spatial and pointwise operations into distinct phases.

A key component of our design is the Adaptive Feature Pyramid Network (AFPN), which replaces computationally expensive deconvolution-based upsampling with an efficient multi-scale feature fusion strategy. By aggregating feature maps at different resolutions, AFPN ensures the robust handling of diverse poses and body sizes without

incurring the high overhead of traditional upsampling layers. Building on the AFPN, we introduce Dual-Gate Context Blocks (DGCBs) to refine global contextual information, which is essential for managing occlusions and cluttered backgrounds. To further enhance feature representation, our approach incorporates Squeeze-and-Excitation (SE) blocks and a Spatial–Channel Refinement Module (SCRM). SE adaptively recalibrates channel-wise feature responses, while SCRM simultaneously optimizes spatial and channel dimensions, amplifying critical cues. This collaboration of multi-scale aggregation, global context gating, and attention-based refinement significantly improves the visibility of obscured or overlapping joints, ultimately producing more accurate and efficient pose estimation. We adopt a coordinate classification approach instead of generating dense heatmaps. Specifically, each joint's feature representation is passed through Multi-Layer Perceptron (MLP) heads that output discrete horizontal and vertical coordinate estimates, alleviating the quantization errors typical of heatmap-based pipelines and removing the memory and computational overhead required for large-scale heatmap generation and post-processing. This design preserves localization precision while reducing both model size and inference latency. Unlike previous coordinate classification approaches such as SimCC and AECA-PRNetCC, our AMFACPose model uniquely combines explicit spatial awareness through CoordConv2d, multi-scale feature fusion via AFPN, and dual-path attention mechanisms, delivering superior accuracy while maintaining low computational cost, making it highly suitable for real-time deployment in resource-constrained environments.

The following three fundamental contributions emerge from this work:

1. We propose a modified ResNet backbone that replaces the standard convolutions with CoordConv2d and DSC, reducing computational overhead while preserving strong feature extraction capabilities. To further elevate feature quality, the backbone incorporates SE blocks and an SCRM, adaptively enhancing critical regions and channels, which is particularly valuable for partially visible or overlapping keypoints.
2. To eliminate costly deconvolution-based upsampling, we introduce an AFPN that efficiently aggregates multi-scale feature maps. Building on the AFPN, DGCBs refine global context, ensuring the robust handling of scale variations, cluttered backgrounds, and complex human poses across varying resolutions.
3. We validate our AMFACPose model on the COCO and CrowdPose datasets, achieving notable improvements in both accuracy and efficiency over the existing methods. Moreover, our model performance on edge devices demonstrates the practicality of these design choices for deployment in diverse, resource-constrained settings.

The remainder of this paper is organized as follows: Section 2 reviews the key developments in HPE, situating our work within the existing literature. Section 3 introduces the proposed AMFACPose framework, detailing each of its core components, including the modified ResNet backbone and the AFPN with DGCBs. Section 4 explains the experimental setup, datasets, and implementation specifics. Section 5 presents our empirical findings, comparing them against SOTA methods on benchmarks such as COCO and CrowdPose. We also discuss the performance of our model on edge devices in Section 6 and provide ablation studies in Section 7 to isolate the contributions of each architectural component. Finally, Section 8 concludes the paper by summarizing our primary insights and suggesting directions for future research in resource-efficient and high-accuracy 2D HPE.

2. Related Work

DL has substantially transformed 2D HPE by automating feature extraction, leading to improvements in both accuracy and computational efficiency. Early research explored regression-based methods for direct keypoint coordinate prediction. Although these approaches initially faced consistency challenges, the Residual Log-likelihood Es-

timation (RLE) [17] achieved good performance comparable to leading heatmap-based techniques. However, these methods continue to face challenges in handling scale variations and occlusions.

A significant development in 2D HPE occurred with the adoption of two-dimensional Gaussian heatmaps for joint localization. Initially transforming the coordinate prediction task into heatmap generation [18], these methods achieved greater stability. Further progress came from architectures like the Stacked Hourglass Network [6], which utilized symmetric encoder–decoder structures with repeated pooling and upsampling to capture multi-scale features. However, heatmap-based methods can suffer high computational costs due to the requirement for dense heatmaps, large upsampling layers, and post-processing operations such as non-maximum suppression. To alleviate these burdens, FasterPose [19] introduced a more streamlined design, while Dense layer and Identity block Parallel Network (IDPNet) [20] implemented lightweight architectural choices targeted toward resource-constrained deployments.

Despite the accuracy benefits of heatmap-based approaches, quantization errors remain a persistent issue. These errors arise from discretizing joint locations onto the heatmap's pixel grid, which can reduce precision as resolution declines. Various minimization techniques have been proposed, including Taylor expansion [21] to refine predictions around the heatmap peak response, and one-dimensional heatmaps [22], which compress spatial dimensionality without sacrificing localization quality. Furthermore, recent work highlights the role of unbiased data processing in reducing systematic bias [23]. Attention mechanisms, whether spatial and channel-based, also help address occlusions and cluttered scenes. Examples include spatially oriented channel attention for better joint discernment [24] and adaptive efficient channel attention for refined feature recalibration [13].

Given the rising demand for real-time and mobile HPE applications, a critical line of research focuses on designing computationally efficient, accurate architectures. While high-resolution representation learning [8] preserves spatial detail throughout the network, it often leads to significant memory overhead. For instance, SimCC [12] reframed HPE to be compatible with both CNN and Transformer architectures, eliminating the need for dense heatmap predictions. Building on this, A. Zakir et al. [25] proposed an efficient bridge attention integration mechanism that enhances feature representation while maintaining computational efficiency.

An important and emerging direction in HPE focuses on confidence score calibration and keypoint visibility estimation for robust occlusion handling. Jiang et al. [26] introduced HPCVNet, which jointly calibrates confidence scores and explicitly classifies keypoint visibility, achieving a mAP of 77.6 on COCO. In contrast, our method adopts an implicit occlusion-handling strategy using attention-driven modules such as AFPN and DGCB. Without requiring auxiliary visibility branches, AMFACPose achieves 76.6 mAP on COCO and demonstrates strong performance on occlusion-heavy benchmarks such as CrowdPose. These strategies reflect a distinct approach to handling partial visibility and overlapping joints in 2D pose estimation.

Building on these developments, we propose AMFACPose, a unified and lightweight pose estimation framework. Unlike heatmap-based pipelines, AMFACPose employs a coordinate classification strategy, avoiding deconvolution layers, dense heatmaps, and post-processing stages. This design achieves a practical balance between high localization accuracy and computational efficiency, making it well suited for real-world applications constrained by latency, memory, and power.

3. AMFACPose

In 2D HPE, the task is to determine the spatial configuration of human body joints, i.e., keypoints, within an RGB image or video frame [27]. Let the pose **P** be represented by N keypoints, each defined by a 2D coordinate (x_n, y_n). For instance, $N = 17$ in the COCO dataset [27]. Formally, for each individual in the input, the goal is to estimate:

$$\mathbf{P} = \{(x_n, y_n)\}_{n=1}^{N}. \tag{1}$$

To address the challenges of occlusion, scale variation, and excessive computational overhead, we propose AMFACPose as illustrated in Figure 1. Our approach uses a ResNet34 backbone [14] that we modified by integrating CoordConv2d [15] and DSC [16], producing an efficient feature extractor that maintains strong spatial representation. Within this backbone, we incorporate attention modules—SE blocks [28] and an SCRM—to highlight keypoints, relevant channels, and local features, ensuring robust performance under partial occlusions or complex scenes. Next, AFPN fuses multi-scale features extracted from the backbone, retaining both global context and fine-grained details. We then refine these fused features using DGCBs, which selectively enhance relevant contextual information while suppressing background noise. Finally, instead of the commonly used heatmap-based and regression-based approachs, AMFACPose utilizes coordinate classification, thereby avoiding the computationally intensive generation of dense heatmaps and simplifying the inference pipeline. The subsequent sections provide an in-depth exploration of each component, illustrating how AMFACPose successfully balances efficiency with accurate and robust keypoint localization.

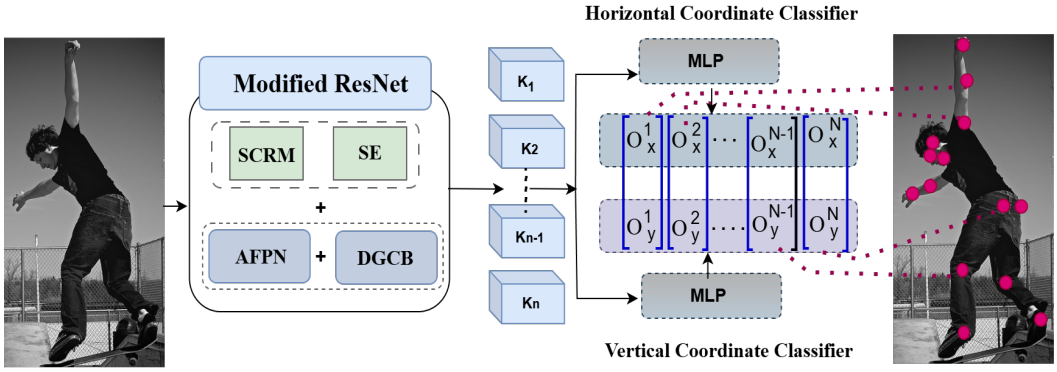

Figure 1. Comprehensive architectural design of AMFACPose for 2D HPE using coordinate classification.

3.1. Modified ResNet Backbone

Recent pose estimation frameworks, such as HRNet [8], Simple Baseline [7], and Stacked Hourglass Networks [6], preserve high-resolution feature representations to achieve precise body–joint localization. Although these high-capacity models attain competitive accuracy, they typically demand substantial computational resources, limiting real-time deployment on edge devices. Vision Transformers [29] likewise exhibit strong performance but often face latency challenges due to expensive self-attention operations.

To balance representational power and computational efficiency, we adopt ResNet34 as our backbone. Compared with deeper variants like ResNet50 or ResNet101, ResNet34 retains the skip connections essential for stable gradient flow [30] yet lowers parameter counts and FLoating-point OPerations (FLOPs). This design offers fine-grained feature extraction necessary for accurate joint detection without incurring prohibitive overhead.

Figure 2 presents an overview of our modified ResNet34 architecture. We remove the final average pooling and fully connected layers, preserving spatial detail in deeper

stages and allowing subtle body part cues to remain accessible. Furthermore, the standard 7×7 input convolution is replaced by a sequence of 3×3 convolutions interleaved with CoordConv2d. Each basic block of the ResNet34 is also enhanced with DSC to reduce computational complexity while maintaining expressive capacity. This modified backbone thus strikes a favorable balance between accuracy and real-time feasibility, serving as the foundation for AMFACPose.

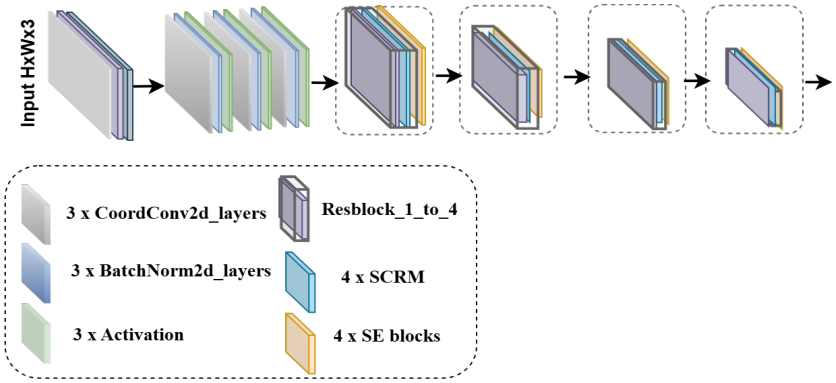

Figure 2. Architecture overview of modified ResNet as feature extractor.

3.1.1. Integration of CoordConv2d for Enhanced Spatial Awareness

The network's initial stem, as shown in Figure 2, incorporates CoordConv2d to embed explicit spatial features at the earliest stage. Let $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$ denote the input tensor, where B is the batch size, C is the channel count, and H, W are spatial dimensions. We first construct normalized coordinate grids $\mathbf{X}_{\text{coord}}$, $\mathbf{Y}_{\text{coord}} \in [-1, 1]^{H \times W}$, providing a consistent reference frame for each pixel location. Four learnable parameters α_x, β_x, α_y, β_y then adaptively scale and shift these coordinates, as follows:

$$\mathbf{X}' = \left(\mathbf{X} \mid \alpha_x \mathbf{X}_{\text{coord}} + \beta_x \mid \alpha_y \mathbf{Y}_{\text{coord}} + \beta_y \right) \in \mathbb{R}^{B \times (C+2) \times H \times W}. \tag{2}$$

After concatenation, a standard 3×3 convolution processes both the original feature maps and these position-aware channels in tandem.

Algorithm 1 summarizes the key steps of CoordConv2d. By retaining explicit spatial information, the stem ensures improved joint localization even under occlusions or viewpoint shifts. The learnable parameters α and β enable flexible adjustment to variations in scale and perspective.

Algorithm 1 Coordinate-enhanced convolution layer

1: **Input:** Feature tensor $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$
2: **Output:** Enhanced feature maps $\widetilde{\mathbf{X}}$
3: **Step 1: Coordinate Grid Initialization**
Initialize normalized grids $\mathbf{X}_{\text{coord}}, \mathbf{Y}_{\text{coord}} \in [-1, 1]^{H \times W}$.
4: **Step 2: Learnable Parameters**
α_x, β_x, α_y, β_y are updated by backpropagation.
5: **Step 3: Coordinate Transformation**
$\mathbf{X}'_{\text{coord}} \leftarrow \alpha_x \mathbf{X}_{\text{coord}} + \beta_x$, $\mathbf{Y}'_{\text{coord}} \leftarrow \alpha_y \mathbf{Y}_{\text{coord}} + \beta_y$.
6: **Step 4: Feature Concatenation**
$\mathbf{X}' \leftarrow [\mathbf{X} \mid \mathbf{X}'_{\text{coord}} \mid \mathbf{Y}'_{\text{coord}}] \in \mathbb{R}^{B \times (C+2) \times H \times W}$.
7: **Step 5: Convolution**
$\widetilde{\mathbf{X}} \leftarrow \text{Conv2D}(\mathbf{X}')$.
8: **Step 6: Output**
return $\widetilde{\mathbf{X}}$.

3.1.2. Depthwise Separable Convolutions for Efficiency

Maintaining spatial detail is crucial for pose estimation, yet computational efficiency is equally important for real-time systems. To address this, each Residual block in ResNet34 replaces the standard 2D convolutions with DSC. In the following equation, a depthwise step applies a unique spatial filter to each input channel:

$$\mathbf{X}_{\text{depthwise}} = \text{Conv2D}_{\text{depthwise}}(\mathbf{X}), \tag{3}$$

where the total parameter count and FLOPs are significantly reduced. A subsequent 1×1 pointwise convolution integrates cross-channel information, as follows:

$$\widetilde{\mathbf{X}} = \text{Conv2D}_{1\times1}(\mathbf{X}_{\text{depthwise}}). \tag{4}$$

Within our Residual blocks, skip connections [14] preserve gradient flow across these separable layers, retaining the ability to learn rich features. When downsampling is needed, e.g., for stride 2, a lightweight residual path aligns the input and output dimensions without adding substantial overhead. Integrating CoordConv2d-based positional encoding and DSC provides a streamlined expressive backbone, striking a strong balance between accuracy and speed. This backbone then serves as the basis for the subsequent modules in AMFACPose.

3.2. Adaptive Feature Pyramid Network (AFPN)

Multi-scale feature fusion is central to effective pose estimation, since body parts can appear at various scales and contextual features may span multiple receptive fields. Traditional approaches, including feature pyramid networks with top–down pathways [31], often rely on deconvolution layers or complex lateral connections, which can amplify computational costs. To address this, we introduce an AFPN that unifies multi-scale representations while preserving critical spatial details, as illustrated in Figure 3.

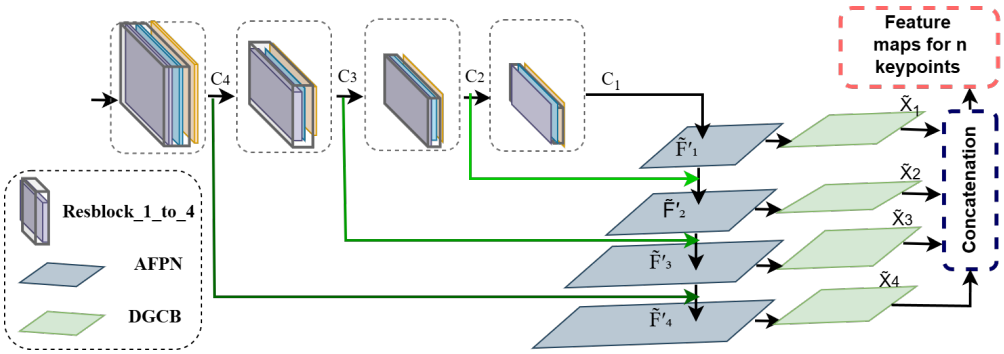

Figure 3. Overview of AFPN architecture. Diagram illustrates four ResNet blocks producing multi-scale feature maps (C_1, C_2, C_3, C_4) that are processed through AFPN and DGCB before concatenation for keypoint feature map generation.

Let $\{C_1, C_2, C_3, C_4\}$ be the feature maps produced by the modified ResNet34 backbone at consecutive stages, as shown in Figure 3, each with a distinct resolution. A 1×1 convolution is applied to each C_i to standardize the channel dimension, as expressed in the following:

$$\mathbf{F_i} = \text{Conv}_{1\times1}(C_i) \quad \text{for } i \in \{1,2,3,4\}. \tag{5}$$

For each feature map $\mathbf{F}_i$, we apply a DGCB (detailed in Section 3.3) to generate a refined feature representation using two learnable masks—a context mask and a gating mask. Both

masks are generated by small convolutional networks with trainable parameters. Formally, the DGCB's refinement is represented as follows:

$$\mathbf{A}_i = \mathrm{DGCB}(\mathbf{F}_i), \tag{6}$$

where the DGCB internally performs an element-wise multiplication between the input feature map and the context and gating masks, as further explained in Section 3.3, Equation (11).

Both $\mathbf{F}_i$ and $\mathbf{A}_i$ are subsequently upsampled to match the spatial resolution of $\mathbf{F}_4$, where the arrow operator $\uparrow(\cdot,\cdot)$ indicates bilinear interpolation, and $\mathrm{size}(F_4)$ represents the height and width of F_4, as shown in the following:

$$\tilde{\mathbf{F}}_i = \uparrow\big(F_i, \mathrm{size}(F_4)\big), \quad \tilde{\mathbf{A}}_i = \uparrow\big(A_i, \mathrm{size}(F_4)\big). \tag{7}$$

Each feature map $\tilde{F}_i$ is then modulated by its corresponding refined representation $\tilde{A}_i$, as follows:

$$\tilde{\mathbf{F}}'_i = \tilde{F}_i \odot \tilde{A}_i, \tag{8}$$

where $\odot$ denotes the element-wise product. The refined outputs from all scales, $\{\tilde{F}'_1, \tilde{F}'_2, \tilde{F}'_3, \tilde{F}'_4\}$, are concatenated and passed through a 1×1 convolution, as follows:

$$\mathbf{X} = \mathrm{Concat}\big(\tilde{F}'_1, \tilde{F}'_2, \tilde{F}'_3, \tilde{F}'_4\big), \quad \mathbf{F}_{\mathrm{out}} = \mathrm{Conv}_{1\times1}(X). \tag{9}$$

The bilinear upsampling in $\uparrow(\cdot,\cdot)$ ensures all features share the same spatial dimensions, enabling their direct combination.

This operation fuses high-level semantics from deeper layers with localized details from shallower ones, producing a consolidated multi-scale feature tensor F_{out} that captures both global context and fine-grained cues.

The AFPN utilizes bilinear interpolation rather than deconvolution to reduce complexity, and it employs our parameterized DGCB to selectively highlight informative features. This design yields a compact architecture well suited for real-time or resource-limited applications. The resulting multi-scale representation serves as a foundation for subsequent pose estimation modules, enabling more accurate keypoint localization under a broad range of poses and imaging conditions.

The use of bilinear interpolation for upsampling in the AFPN is guided by both theoretical rationale and empirical effectiveness. Bilinear interpolation provides a computationally efficient, parameter-free method for resizing feature maps, making it particularly attractive for real-time or resource-constrained scenarios. In contrast to deconvolution, which increases model complexity and may introduce checkerboard artifacts, bilinear interpolation performs deterministic, smooth upsampling without additional learnable weights.

In our design, the potential limitations of bilinear interpolation, such as information loss or aliasing, are addressed through two mechanisms. First, a 1×1 convolution (Equation (5)) is applied before upsampling, which helps to suppress high-frequency noise and standardize channel dimensions. Second, the upsampled features are modulated by context-aware masks generated from DGCBs (Equation (8)), which selectively enhance salient regions and suppress irrelevant or noisy activation. This combination preserves critical spatial cues while maintaining efficiency.

Unlike traditional Feature Pyramid Networks that rely on fixed top–down pathways for multi-scale feature fusion, the AFPN introduces several architectural enhancements for more effective scale handling beyond the computational advantages of bilinear interpolation. Conventional FPNs typically apply direct addition or fixed-weight fusion, which may overlook the relative importance of scale-specific features. In contrast, the

AFPN integrates Dual-Gate Context Blocks that generate adaptive, context-aware masks to selectively emphasize the most informative features at each scale. Additionally, while traditional FPNs often fuse features sequentially, risking the dilution of fine-grained details from lower levels, the AFPN employs parallel aggregation followed by concatenation and fusion, preserving resolution-specific information. These distinctions collectively enable the AFPN to maintain strong keypoint localization performance across a range of object sizes and challenging visual conditions, while remaining efficient enough for deployment in resource-constrained environments.

3.3. Dual-Gate Context Blocks (DGCBs)

HPE frequently encounters ambiguities stemming from partial occlusions, multiple overlapping individuals, and cluttered scenes. Although previous strategies introduce GCBs or similar modules to incorporate scene-level features [32,33], most rely on a single attention mechanism that may not sufficiently separate background noise from body–joint features. In contrast, our proposed DGCB module learns two distinct masks, a context mask and a gating mask, that work together to refine feature representations and enhance joint localization.

Let $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ be the feature map at a particular scale. A Global Average Pooling (GAP) operation condenses this tensor to $\mathrm{GAP}(\mathbf{X}) \in \mathbb{R}^{B\times C\times 1\times 1}$. Two parallel sets of 1×1 convolutions with ReLU and sigmoid activations process this pooled descriptor, as follows:

$$\mathbf{M}^c(\mathbf{X}) = \sigma\big(\mathbf{W}_2^c \delta\big(\mathbf{W}_1^c \mathrm{GAP}(\mathbf{X})\big)\big), \quad \mathbf{M}^g(\mathbf{X}) = \sigma\big(\mathbf{W}_2^g \delta\big(\mathbf{W}_1^g \mathrm{GAP}(\mathbf{X})\big)\big), \tag{10}$$

where $\delta(\cdot)$ and $\sigma(\cdot)$ denote ReLU and sigmoid activations, and $\mathbf{W}_1^c, \mathbf{W}_2^c, \mathbf{W}_1^g, \mathbf{W}_2^g$ are separate trainable weights, where c and g represent the contextual and gated information.

Although both branches appear structurally similar, they learn to serve distinct functional purposes through several mechanisms. First, the weight parameters are initialized independently and updated separately during training, allowing them to evolve toward different feature spaces. Second, the multiplicative interaction in the final output creates complementary specialization between branches; the network benefits when each mask focuses on different aspects of the input features rather than learning redundant information. Through this design, the context branch captures high-level global semantics features, while the gating branch selectively filters these contextual features based on local activation relevance.

Both masks are broadcast to the shape $\mathbb{R}^{B\times C\times H\times W}$ and applied element-wise to the input, as expressed in the following equation:

$$\widetilde{\mathbf{X}} = \mathbf{X} \odot M^c(\mathbf{X}) \odot M^g(\mathbf{X}), \tag{11}$$

where $\odot$ denotes element-wise multiplication.

Each scale in the AFPN incorporates a DGCB to refine features prior to the final fusion step. Although DGCBs add only two small 1×1 convolutions per scale, they minimally increase computational overhead while substantially boosting keypoint visibility under occlusions or complex backgrounds. Their design segregates global scene information from localized gating cues, fostering more resilient pose estimation in cluttered or overlapping scenarios. A schematic of the DGCB's internal architecture, illustrating the context mask and gating mask flow, is provided in Figure 4.

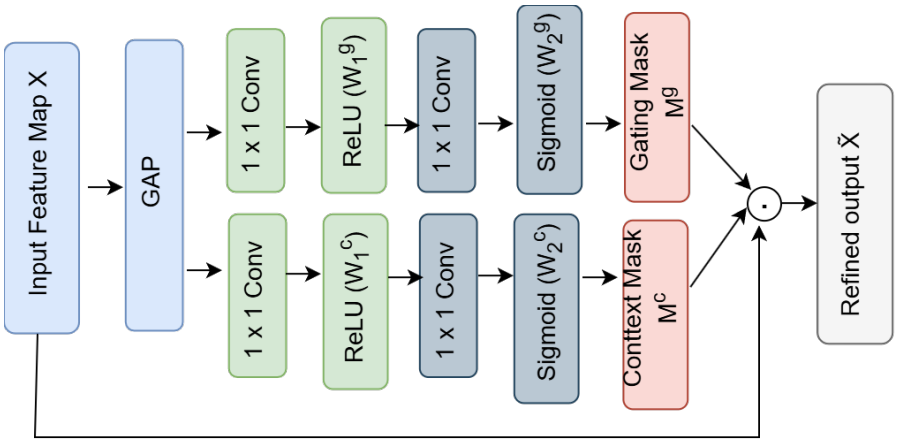

Figure 4. Architecture of DGCB. Two parallel sets of 1×1 convolution paths produce a context mask and a gating mask, each conditioned on the GAP features. These masks are broadcast and multiplied element-wise with the original input, enhancing relevant body–joint features while suppressing background interference.

3.4. Attention Mechanisms: Squeeze-and-Excitation and Spatial–Channel Refinement

Channel- and spatial-level attention can enhance the discriminative power of convolutional backbones for HPE. The proposed architecture incorporates the following two supportive modules: the classical SE block [28], which recalibrates feature channels globally, and a novel SCRM, which jointly emphasizes both channel and spatial dimensions. Although both modules utilize channel-wise weighting, their mechanisms differ fundamentally. SE blocks use a fully-connected bottleneck structure to capture global inter-channel dependencies, while the SCRM integrates a lightweight convolutional transformation for channel attention directly fused with spatial attention. These differences result in collaborative behavior during learning and inference.

3.4.1. Squeeze-and-Excitation (SE) Blocks

Each SE block adaptively re-weights channels based on a global context vector [28], effectively amplifying body part features and suppressing less relevant activations. Formally, let $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ be the incoming feature map. GAP yields $\mathbf{z} = \text{pool}(\mathbf{X}) \in \mathbb{R}^{B\times C\times 1\times 1}$. A small fully connected (FC) network equivalent to 1×1 convolutions, equipped with ReLU and sigmoid activations, has parameters $\mathbf{W}_1, \mathbf{W}_2$, as expressed in the following:

$$\mathbf{s} = \sigma\big(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})\big), \tag{12}$$

where $\delta(\cdot)$ and $\sigma(\cdot)$ denote the ReLU and sigmoid functions, respectively. The resulting channel attention vector $\mathbf{s} \in \mathbb{R}^{B\times C\times 1\times 1}$ is broadcast and multiplied element-wise with $\mathbf{X}$, as follows:

$$\widetilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{s}. \tag{13}$$

Integrating an SE block after each of the ResNet34 backbone's four residual blocks ensures that channel refinements benefit from multi-scale global context. Figure 5 visually summarizes the main steps of the SE module.

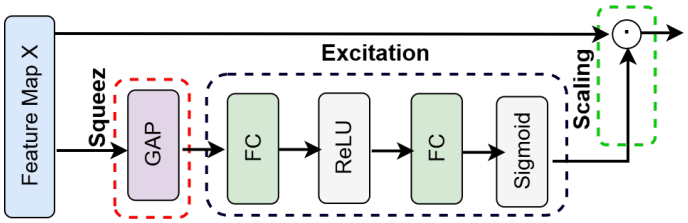

Figure 5. Architecture of SE Module. SE block applies GAP to each channel, producing a single descriptor vector. Two FC layers with ReLU and sigmoid activations re-weight channels based on their global importance, allowing network to emphasize key body part features.

3.4.2. Spatial–Channel Refinement Module (SCRM)

While SE blocks excel at global-channel-level recalibration, local spatial dependencies are crucial for accurately locating body joints, especially under occlusions. Addressing this need, our novel SCRM provides a simultaneous refinement of both the spatial and channel dimensions through a dual-attention mechanism. Let $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ be the input feature map. An Adaptive Average Pooling (AAP) operation condenses $\mathbf{X}$ to $\mathrm{AAP}(\mathbf{X}) \in \mathbb{R}^{B\times C\times 1\times 1}$, which is then transformed into a channel attention vector by a 1D convolution, as shown in the following:

$$\mathbf{X_c} = \sigma\big(\mathrm{BN}(\mathrm{Conv1D}(\mathrm{pool}(\mathbf{X})))\big), \tag{14}$$

where $\sigma(\cdot)$ and $\mathrm{BN}(\cdot)$ represent the sigmoid and BN functions, respectively. In parallel, a depthwise 3×3 convolution with BN and sigmoid activation extracts a spatial attention mask, as follows:

$$\mathbf{X_s} = \sigma\big(\mathrm{BN}(\mathrm{Conv2D}_{\mathrm{depthwise}}(\mathbf{X}))\big). \tag{15}$$

Broadcasting both $\mathbf{X_c}$ and $\mathbf{X_s}$ to $\mathbb{R}^{B\times C\times H\times W}$ and multiplying them element-wise with $\mathbf{X}$ yields the following:

$$\widetilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{X_c} \odot \mathbf{X_s}. \tag{16}$$

Although channel attention in SCRM is similar to SE, several critical distinctions differentiate their functionalities and effects. First, the channel attention generation differs architecturally. SE uses a bottleneck FC structure, compressing channel dimensions and modeling global inter-channel dependencies in a non-linear transformation, thereby capturing abstract relationships among the channels. In contrast, SCRM maintains the original channel dimensionality via 1D convolution, preserving channel-specific information without dimensionality reduction, thereby modeling simpler yet spatially aware channel relationships.

Second, the most significant distinction lies in the SCRM's simultaneous integration of spatial attention, which SE blocks lack entirely. By coupling channel emphasis ($\mathbf{X_c}$) with spatial filtering ($\mathbf{X_s}$), the SCRM enables the network to focus on which feature channels matter and where within the spatial field they should be emphasized. This dual optimization is particularly beneficial for the accurate localization of joints under challenging conditions such as partial occlusions or complex human poses. Unlike sequential attention methods such as CBAM [33], which apply spatial and channel attention independently in sequence, the SCRM fuses them in a single step for improved efficiency. Thus, the global channel recalibration of SE blocks and the combined spatial–channel refinement of the SCRM provide attention functionalities and collectively improve the robustness and accuracy of the model. Figure 6 outlines the SCRM's structure and highlights its combined spatial–channel approach.

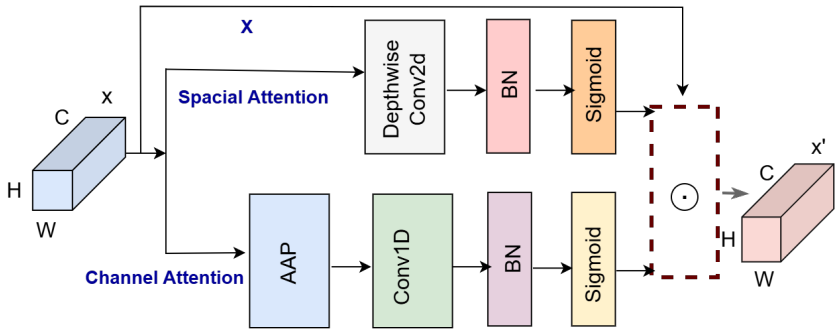

Figure 6. Overview of SCRM structure. A 1D convolutional layer generates channel attention from globally pooled features, while a depthwise 3×3 convolution extracts spatial attention. Multiplying both attention maps into the original feature refines local body part features and strengthens channel emphasis, helping to resolve occlusions or overlapping subjects.

3.5. AMFACPose Head and Coordinate Classification

Most HPE frameworks generate heatmaps for each joint and use peak-finding or regression to extract (x, y) coordinates, which can inflate memory usage and computational complexity [7,8,24]. In AMFACPose, we replace heatmaps with a coordinate classification approach that directly predicts discrete (x, y) indices for each joint. This methodology reduces the overhead of creating high-resolution heatmaps, enabling efficient inference without compromising accuracy.

3.5.1. Final Feature Reorganization

After passing through the modified ResNet34 backbone, AFPN, and attention modules, i.e., DGCBs, SE, and SCRM, the network outputs a fused feature map $\mathbf{F} \in \mathbb{R}^{B \times C \times H' \times W'}$, where B is the batch size, C is the number of channels, and (H', W') denotes the reduced spatial dimensions. To enable per-joint classification, the channel dimension C is reshaped into $(\mathrm{N} \times d)$, where d represents an embedding size for each joint, as expressed in the following:

$$\mathbf{F}' = \text{reshape}\Big(\mathbf{F}, B, \mathrm{N}, d, (H' \times W')\Big). \tag{17}$$

This transformation allocates a dedicated d-dimensional embedding for every joint at each spatial location, ensuring that subsequent classifiers can learn rich features specific to each keypoint.

3.5.2. Discrete Coordinate Classification

We discretize the continuous input space (W, H) into N_x and N_y bins along the horizontal and vertical axes, respectively, as follows:

$$N_x = W \times k, \quad N_y = H \times k, \tag{18}$$

where $k \geq 1$ is a scaling factor that controls the granularity of coordinate discretization. Each ground-truth joint coordinate (x_i, y_i) is mapped to a discrete location in $\{1, \ldots, N_x\} \times \{1, \ldots, N_y\}$.

In our implementation, we adopt $k = 2$, following the design choice proposed by Li et al. [12]. This value achieves a strong balance between prediction granularity and computational efficiency. Larger values of k lead to finer bins, which increase memory and computation requirements with limited benefit, while smaller values reduce model cost but introduce quantization artifacts that degrade localization precision.

For each joint i, the reorganized feature $\mathbf{F}'_i \in \mathbb{R}^{B\times d\times(H'\times W')}$ is fed into two MLPs equipped with Mish activation functions, as follows:

$$\mathbf{p}_x^i = \mathrm{MLP}_x\left(\mathbf{F}'_i\right) \in \mathbb{R}^{N_x}, \quad \mathbf{p}_y^i = \mathrm{MLP}_y\left(\mathbf{F}'_i\right) \in \mathbb{R}^{N_y}. \tag{19}$$

where $\mathbf{p}_x^i$ and $\mathbf{p}_y^i$ represent discrete probability distributions over the set of possible x- and y-bins. Algorithm 2 outlines this procedure.

Algorithm 2 AMFACPose: keypoint estimation process

Require: RGB image I of size $H \times W \times 3$
Ensure: Predicted keypoint coordinates $\{o_x^1, o_y^1, \ldots, o_x^N, o_y^N\}$ for N keypoints
Feature Extraction:
  Process I through the modified ResNet backbone to obtain fused feature map $\mathbf{F}$ of size (B, C, H', W')
Feature Reorganization:
  Reshape $\mathbf{F}$ to $\mathbf{F}'$ of size $\left(B, \mathrm{N}, d,\ H' \times W'\right)$
Discretization Setup:
  Let $k \geq 1$ be the scaling factor
  Set $N_x = W \times k$ and $N_y = H \times k$
for $i = 1, \ldots, \mathrm{N}$ **do**
    Horizontal Classification:
      $\mathbf{p}_x^i \leftarrow \mathrm{MLP}_x\left(\mathbf{F}'_i\right)$
    Vertical Classification:
      $\mathbf{p}_x^i \leftarrow \mathrm{MLP}_y\left(\mathbf{F}'_i\right)$
end for
return Keypoint coordinates $\{o_x^1, o_y^1, \ldots, o_x^N, o_y^N\}$

By framing joint location prediction as a classification problem, AMFACPose streamlines the output space, bypasses large heatmaps, and simplifies post-processing. The MLP-based classifiers with Mish activations can learn complex spatial dependencies, ultimately leading to improved localization precision. This framework also reduce memory usage, making the model more suitable for real-time and resource-constrained scenarios.

3.6. AMFACPose Loss Function: KLDDiscretLoss

Conventional HPE often adopts Mean Squared Error (MSE) [7] or L1-based losses, which assume continuous error distributions [34]. However, in our coordinate classification framework, joint positions are discretized into bins along the x- and y-axes, rendering such regression-focused objectives less optimal. To address this discrepancy, we propose KLDDiscretLoss, a divergence-based criterion grounded in Kullback–Leibler Divergence (KLD) [35]. By treating pose estimation as a classification problem, KLDDiscretLoss directly compares predicted probability distributions with discrete ground-truth distributions, thereby capturing the inherent uncertainties of joint positions.

A key advantage of KLDDiscretLoss is that it models probability distributions rather than point estimates. This perspective is particularly beneficial when joint locations are ambiguous due to scale variations, occlusions, or overlapping body parts. In order to refine the network's confidence calibration, we incorporate two additional mechanisms—label smoothing [36] and temperature scaling [37]. Label smoothing allocates a small fraction of the ground-truth probability mass uniformly across all coordinate bins, preventing overfitting and minimizing cases where the network becomes overconfident in a single discrete location.

Temperature scaling, controlled by a parameter T, modifies the softmax logits by $\frac{1}{T}$. When $T > 1$, the resulting distributions become softer, reflecting higher uncertainty in the model's predictions; when $T < 1$, the distributions sharpen, forcing the network to

commit more strongly to specific bins. The temperature value plays an important role in situations involving occlusion or pose ambiguity. A moderately sharpened output distribution encourages the model to focus on likely joint locations, improving spatial localization while still expressing uncertainty. In our experiments, we set $T = 0.8$, which we found to provide a favorable trade-off between sharpness and calibration. This choice was guided by early empirical validation and is consistent with insights from a previous study on model calibration [37]. It allows the network to remain confident in its predictions without becoming overly rigid or under-responsive in uncertain contexts, such as occluded joints.

Concretely, let o_x^i and o_y^i denote the predicted logits for x- and y-coordinates of the i-th joint, and let $\text{gt}(o_x^i)$ and $\text{gt}(o_y^i)$ be the corresponding ground-truth distributions. A joint-specific weight W_i is assigned to emphasize harder-to-detect keypoints, such as hands or feet. The KLDDiscretLoss for the i-th joint is given by the following:

$$\text{Loss}_\text{i} = W_i \times \left[\text{KLD}\left(\log\left[\text{Softmax}\left(\frac{o_x^i}{T} \right) \right], \text{gt}(o_x^i) \right) + \text{KLD}\left(\log\left[\text{Softmax}\left(\frac{o_x^i}{T} \right) \right], \text{gt}(o_x^i) \right) \right], \tag{20}$$

where $\text{Softmax}(\frac{o_x^i}{T})$ and $\text{Softmax}(\frac{o_y^i}{T})$ convert the scaled logits into probability distributions. The total KLDiscretLoss is then computed as the mean across all N joints, as follows:

$$\text{KLDDiscretLoss} = \frac{1}{N} \sum_{i=1}^{N} \text{Loss}_i. \tag{21}$$

Compared to regression-based losses such as SmoothL1 or MSE, KLDDiscretLoss offers a principled advantage under occlusion. Regression losses penalize deviations from ground-truth coordinates without accounting for uncertainty, which can result in overconfident and unreliable predictions for occluded or ambiguous joints. In contrast, KLDDiscretLoss allows the network to express uncertainty by distributing probability mass across plausible locations. This soft probabilistic output creates natural error bounds. If a joint is fully occluded, the prediction can approach a uniform distribution, with the error exceeding the expected value by up to half the discretization range. Empirically, this advantage is reflected in our CrowdPose performance, where AMFACPose achieves 65.9 AP in the hard subset, demonstrating robustness in severe occlusion scenarios. Thus, KLDDiscretLoss provides a more reliable and uncertainty-aware mechanism for keypoint localization than point-based regression losses.

Our PyTorch-based implementation [38] processes the (x, y) distributions for each joint independently, facilitating the straightforward integration of label smoothing and temperature scaling in the preprocessing steps. This structure also enables fine-grained control over which joints receive higher weighting, enabling the model to spend more capacity on challenging joints or underrepresented body parts. By guiding the network to produce calibrated probability distributions rather than single-point predictions, KLDiscretLoss enhances robustness against partial visibility, background clutter, and pose variability.

4. Experimental Setup

4.1. Datasets

We conducted comprehensive evaluations of AMFACPose using two established benchmarks in HPE—the MS COCO [27] and CrowdPose datasets [39]. These datasets were selected for their complementary characteristics, enabling a thorough assessment of our model across diverse scenarios.

The MS COCO 2017 dataset serves as a primary benchmark for HPE evaluation, containing over 200,000 images with approximately 250,000 annotated person instances.

Each instance is labeled with 17 keypoints, encompassing facial features i.e., eyes, ears, nose, and body joints such as shoulders, elbows, wrists, hips, knees, ankles. The dataset is partitioned into 118,000 training images, 5000 validation images, and a separate test set. MS COCO's strength lies in its diversity, featuring varied poses, scales, and occlusions in natural contexts, thereby providing a robust evaluation framework for model generalization.

On the other hand, the CrowdPose dataset [39] specifically addresses the challenges of pose estimation in crowded scenarios. Comprising 20,000 images with approximately 80,000 person instances, CrowdPose annotates 14 keypoints per person, focusing on body joints i.e., shoulders, elbows, wrists, hips, knees, ankles, while excluding facial landmarks. The dataset is divided into 10,000 training, 2000 validation, and 8000 testing sets. CrowdPose's distinctive feature is its emphasis on person-to-person occlusions and high-density scenarios, presenting more challenging conditions than the typical pose estimation datasets. This characteristic makes it particularly valuable for evaluating our model's performance in real-world crowded environments, where accurate pose estimation is crucial yet technically challenging.

4.2. Evaluation Metrics

Our model's performance evaluation utilizes the Object Keypoint Similarity (OKS) metric, which provides a rigorous assessment of keypoint localization accuracy. The OKS metric quantifies the similarity between predicted and ground-truth keypoint positions through the following formulation:

$$\text{OKS} = \frac{\sum_i \delta(v_i > 0) \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)}{\sum_i \delta(v_i > 0)} \tag{22}$$

where d_i represents the Euclidean distance between the predicted and ground-truth positions for the i-th keypoint, s denotes the person instance scale, k_i is a keypoint-specific normalization constant, and v_i indicates keypoint visibility. The indicator function $\delta(v_i > 0)$ ensures evaluation focuses exclusively on visible keypoints. We used average precision (AP) as our primary performance metric, calculated across ten OKS thresholds ranging from 0.50 to 0.95 in 0.05 increments. This comprehensive range allows for a detailed performance assessment at various precision levels. We specifically report AP_{50} and AP_{75} corresponding to OKS thresholds of 0.50 and 0.75, providing insights into the model's performance at different precision requirements. Scale-specific metrics AP_M and AP_L evaluate performance on medium- and large-sized instances, respectively. Additionally, we compute average recall (AR) following similar protocols to AP, offering complementary insights into the model's detection capabilities.

For the CrowdPose dataset evaluation, we maintain consistency with MS COCO by utilizing the same fundamental OKS metric while incorporating additional crowd-specific measures. These include AP metrics stratified by scene complexity, AP_{easy} , AP_{medium}, and AP_{hard}. Scene complexity classification is determined by the crowding level, computed as the average Intersection over Union of the ground-truth bounding boxes within each image. This stratified evaluation framework enables a detailed assessment of our model's performance across varying levels of scene complexity and person-to-person occlusion. Through this comprehensive evaluation framework, combining standard OKS-based metrics with crowd-specific measures, we ensure a thorough assessment of our model's keypoint localization capabilities across diverse scenarios. This approach validates the model's reliability in both general and crowded environments, providing a complete understanding of its real-world applicability.

4.3. Implementation Details

Our implementation strategy emphasizes training stability, efficient resource utilization, and strong generalization for real-world pose estimation. We implemented comprehensive data augmentation techniques including random horizontal flips, rotational variations from $-30°$ to $+30°$, and scale adjustments from 0.7 to 1.3 [40]. These augmentations were implemented using the PyTorch 1.12.1 framework, ensuring efficient and reliable model training.

Training proceeds for 140 epochs, with a batch size of 32 to maintain a balance between gradient stability and computational throughput. Six parallel data-processing workers further accelerate input pipelines. The initial learning rate is set to 1×10^{-5}, enabling gradual convergence while preventing overshooting of local minima. We used the Mish activation function [41] in the ResNet, a smooth and non-monotonic alternative to ReLU that has demonstrated effectiveness in minimizing vanishing gradients [42] and improving feature extraction in deeper models.

To refine parameter updates and manage regularization, we adopt the AdamW optimizer [43], which decouples weight decay from the main optimization steps. This separation grants more precise control over the magnitude of regularization and often produces better generalization performance [44]. Empirical studies [45] show that AdamW outperforms classical optimizers in complex tasks by maintaining stable gradients and resisting overfitting, making it particularly suitable for the challenges posed by dense keypoint localization.

The full AMFACPose model—which includes the AFPN, DGCBs, and other attention modules—requires approximately 78 h to converge, while the baseline ResNet34 model converges in approximately 42 h under the same training schedule. Despite the modest increase in training time, the additional modules yield significant accuracy gains, justifying their computational cost.

5. Results and Discussion

This section presents a comprehensive quantitative and qualitative evaluation of the proposed AMFACPose framework. We first detail its performance on the MS COCO dataset, highlighting both accuracy and scalability. Subsequently, we examine resource–efficiency trade-offs and analyze the model's behavior under congested scenarios using the CrowdPose dataset. Finally, we provide qualitative examples of the model predictions, illustrating AMFACPose's versatility across diverse real-world conditions.

5.1. Performance on COCO Dataset

Table 1 compares AMFACPose with several SOTA 2D HPE models on the MS COCO dataset, including multiple recent approaches. Utilizing a ResNet34 backbone with an input resolution of 384×288, AMFACPose achieves an AP of 76.6, surpassing coordinate-based methods such as AECA-PRNetCC, with an AP of 76.0, and SimCC, with an AP of 73.4. Additionally, AMFACPose marginally outperforms the strong heatmap-based baseline HRNet-W48, which achieves an AP of 76.3. The method also demonstrates superior performance compared to recently introduced techniques, such as BR-Pose with an AP of 75.3, various PCDPose models exhibiting AP scores ranging from 73.5 to 74.3, SDPose variants achieving AP scores between 73.5 and 73.7, and the CSDNet-m/12 model with an AP of 75.0. Many of these methods rely on the robust HRNet backbone, emphasizing the competitive advantage of AMFACPose's coordinate classification pipeline, which achieves comparable or superior accuracy without incurring significant computational overhead from heatmap generation and post-processing.

Table 1. Quantitative comparison on MS COCO; accuracy metrics across different model architectures and input configurations.

Model	Backbone	Input	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR
Heatmap-based								
SimpleBaseline [7]	ResNet-50	384 × 288	72.2	89.3	78.9	68.1	79.7	77.6
	ResNet-101	384 × 288	73.6	89.6	80.3	69.9	81.1	79.1
HRNet [8]	HRNet-W32	384 × 288	75.8	90.6	82.7	71.9	82.8	81.0
	HRNet-W48	384 × 288	76.3	90.8	82.9	72.3	**83.4**	**81.2**
TokenPose-L/D24 [10]	CNN	256 × 192	75.8	90.3	82.5	72.3	82.7	80.6
TokenPose-L/D6 [10]	CNN	256 × 192	75.4	90.0	81.8	71.8	82.4	80.4
BR-Pose [46]	HRNet-W32	256 × 192	75.3	90.6	82.5	71.7	81.9	80.4
PCDPose-B [47]	HRNet-W32	256 × 192	74.3	89.7	81.4	70.8	81.0	79.5
PCDPose-S-V2 [47]	HRNet-W32	256 × 192	74.1	89.5	81.1	70.7	81.0	79.3
PCDPose-S-V1 [47]	HRNet-W32	256 × 192	73.5	89.6	80.9	69.8	80.9	78.9
SDPose-B [48]	CNN	256 × 192	73.7	89.6	80.4	70.3	80.5	79.1
SDPose-S-V2 [48]	CNN	256 × 192	73.5	89.5	80.4	70.1	80.3	78.7
CSDNet-m/12 [49]	HRNet-W32	256 × 192	75.0	89.9	81.7	71.4	81.9	80.1
Regression-based								
PRTR [50]	ResNet-50	384 × 288	68.2	88.2	75.2	63.2	76.2	76.0
	ResNet-101	384 × 288	70.1	88.8	77.6	65.7	77.4	77.5
	HRNet-W32	384 × 288	73.1	89.4	79.8	68.8	80.4	79.8
Coordinate-based								
SimCC [12]	ResNet-50	384 × 288	73.4	89.2	80.0	69.7	80.6	78.8
AECA-PRNetCC [13]	ResNet34	384 × 288	76.0	92.5	82.4	73.3	80.7	79.0
AMFACPose	**ResNet34**	**384 × 288**	**76.6**	**92.6**	**83.7**	**73.9**	81.2	79.3
	ResNet34	256 × 256	75.6	**92.6**	81.8	72.5	80.2	78.2
	ResNet18	384 × 288	73.1	91.6	79.5	70.3	78.0	76.0
	ResNet18	256 × 256	72.1	91.5	79.4	69.1	76.7	75.0

A detailed analysis of the threshold-specific metrics reveals strong performances at both moderate and stricter accuracy levels. Specifically, AMFACPose achieves an AP_{50} of 92.6 and an AP_{75} of 83.7. Additionally, the method demonstrates balanced effectiveness across different object scales, achieving an AP_M of 73.9 and an AP_L of 81.2. These results show the effectiveness of the Adaptive Feature Pyramid Network and the attention modules in managing subjects of varying sizes and enhancing global and local contextual understanding, even under challenging conditions such as partial occlusions or diverse poses. To further investigate the trade-offs between resource efficiency and accuracy, evaluations with smaller input resolutions, such as 256 × 256, and lighter backbones, such as ResNet18, were conducted. These configurations consistently maintain AP scores above 72.0, demonstrating AMFACPose's adaptability to varying computational constraints. Notably, the most compact configuration with ResNet18 at a resolution of 256 × 256 still achieves an AP of 72.1, competitively close to several recent, larger-architecture methods. This adaptability highlight the practical applicability of AMFACPose, particularly for deployment in real-world scenarios involving edge devices with limited computational resources.

5.2. Model Complexity and Resource Efficiency

Table 2 presents a comparative summary of AMFACPose alongside recent 2D HPE models, emphasizing accuracy and computational efficiency. With a ResNet34 backbone at 384 × 288 input, AMFACPose achieves an AP of 76.6 using only 3.8 M parameters and 5.2 GFLOPs. This reflects a substantial improvement over HRNet-W48, which reports a slightly lower AP of 76.3 but requires 63.6 M parameters and 32.9 GFLOPs. Compared to

AECA-PRNetCC, which achieves an AP of 76.0 with 29.0 M parameters and 8.3 GFLOPs, AMFACPose delivers similar accuracy at a fraction of the computational cost.

Table 2. Model complexity on MS COCO: AP, parameters (M), and FLOPs (G) for leading 2D HPE methods.

Model	Backbone	Input	AP	Params (M)	FLOPs (G)
Heatmap-based					
SimpleBaseline [7]	ResNet-50	384 × 288	72.2	34.0	18.6
	ResNet-101	384 × 288	73.6	53.0	26.7
HRNet [8]	HRNet-W32	384 × 288	75.8	28.5	16.0
	HRNet-W48	384 × 288	76.3	63.6	32.9
TokenPose-L/D24 [10]	CNN	256 × 192	75.8	27.5	11.0
TokenPose-L/D6 [10]	CNN	256 × 192	75.4	20.8	9.1
BR-Pose [46]	HRNet-W32	256 × 192	75.3	31.3	9.0
PCDPose-B [47]	HRNet-W32	256 × 192	74.3	13.8	5.2
PCDPose-S-V2 [47]	HRNet-W32	256 × 192	74.1	7.7	6.7
PCDPose-S-V1 [47]	HRNet-W32	256 × 192	73.5	8.0	4.5
SDPose-B [48]	CNN	256 × 192	73.7	13.2	5.2
SDPose-S-V2 [48]	CNN	256 × 192	73.5	6.2	4.7
CSDNet-m/12 [49]	HRNet-W32	256 × 192	75.0	17.4	6.9
Regression-based					
PRTR [50]	ResNet-50	384 × 288	68.2	41.5	11.0
	ResNet-101	384 × 288	70.1	60.4	19.1
	HRNet-W32	384 × 288	73.1	57.2	21.6
Coordinate-based					
SimCC [12]	ResNet-50	384 × 288	73.4	36.8	20.2
AECA-PRNetCC [13]	ResNet34	384 × 288	76.0	29.0	8.3
AMFACPose	ResNet34	384 × 288	76.6	3.8	5.2
	ResNet34	256 × 256	75.6	3.6	3.1
	ResNet18	384 × 288	73.1	2.5	3.2
	ResNet18	256 × 256	72.1	2.4	1.9

Recent models further highlight AMFACPose's efficiency. BR-Pose reaches 75.3 AP with 31.3 M parameters and 9.0 GFLOPs. PCDPose variants yield AP scores from 73.5 to 74.3, with 7.7–13.8 M parameters and 4.5–6.7 GFLOPs. SDPose methods achieve 73.5–73.7 AP with 6.2–13.2 M parameters and 4.7–5.2 GFLOPs, while CSDNet-m/12 reaches 75.0 AP with 17.4 M parameters and 6.9 GFLOPs. In all cases, AMFACPose offers better accuracy with significantly lower resource requirements, supporting its deployment in constrained environments.

Further reductions are achieved using ResNet18. At 256 × 256 input, AMFACPose maintains 72.1 AP with only 2.4 M parameters and 1.9 GFLOPs, showing its adaptability to low-power applications without severe performance loss.

We also visualize these trade-offs in Figure 7, which plots the AP on the vertical axis versus model parameters on the horizontal axis. The size and color of each bubble correspond to FLOPs, and the outline color denotes the underlying methodology, which are heatmap-based, regression-based, and coordinate-based. As shown, AMFACPose configurations appear near the lower-parameter, lower-FLOP regions while achieving competitive accuracy. In contrast, HRNet-W48 occupies a higher-parameter area with significantly higher computational cost for a similar AP score. These findings point to the efficacy of AMFACPose's coordinate classification pipeline and lightweight design components, making it well suited for resource-constrained scenarios.

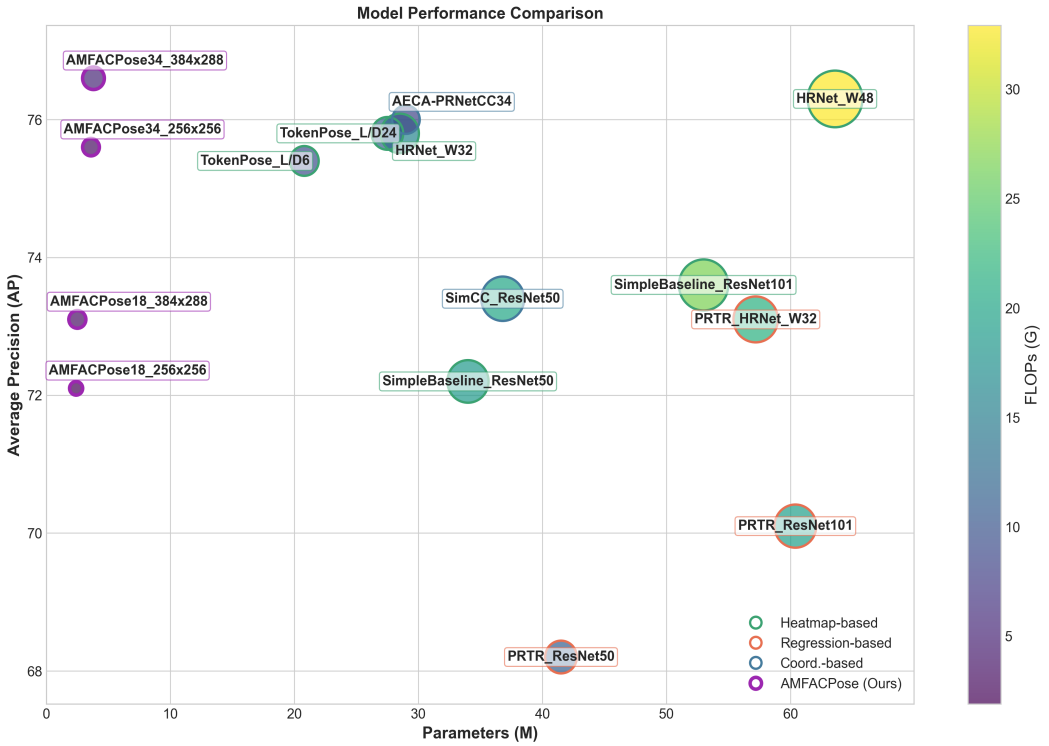

Figure 7. Comparative visualization of model efficiency. Bubbles represent both size and color of FLOPs, while *x*-axis shows parameter count, and *y*-axis depicts AP. AMFACPose with purple outline maintains high AP with fewer parameters and FLOPs compared to various SOTA models.

While demonstrating significant improvements in computational efficiency, AMFACPose also introduces several practical trade-offs. The use of DSC and the AFPN substantially reduces computation to 5.2 GFLOPs, compared to 32.9 GFLOPs in HRNet-W48, while maintaining competitive AP scores. This design enables real-time performance even on edge devices such as Jetson platforms, as discussed in Section 6. However, the coordinate classification strategy, while efficient, involves discretizing the coordinate space, which may introduce localization limitations for small joints (e.g., wrists or ankles) when high precision is required. In such cases, high-resolution heatmap regression may provide better granularity. Moreover, while our approach handles moderate occlusions effectively using attention modules such as DGCBs and the SCRM, it does not incorporate explicit visibility classification. Competing works such as HPCVNet [26], which achieves 77.6 mAP on COCO, model keypoint visibility directly, offering added robustness in scenarios with extreme occlusion or dense overlapping subjects. These trade-offs reflect the broader goal of balancing accuracy, interpretability, and deployment feasibility in real-world pose estimation systems.

5.3. Performance on CrowdPose Dataset

We further evaluated AMFACPose on the CrowdPose dataset [39], known for emphasizing challenging scenarios involving person-to-person occlusions and complex group interactions. As detailed in Table 3, AMFACPose utilizing a ResNet34 backbone achieves a leading AP score of 75.3, along with AP_{50} and AP_{75} values of 93.4 and 81.0, respectively. This performance surpasses established frameworks such as PRTR with an AP of 71.6, HRFormer with 72.6, and ED-Pose with Swin-L, which achieves 73.1. Furthermore, AMFACPose demonstrates higher accuracy than recent methods, including GroupPose with an AP of 74.1, MAQT with 74.3, and CCAM-Person with 74.4.

Across varying levels of difficulty, AMFACPose maintains strong performance, reporting an AP_{easy} of 82.1 and AP_{medium} of 76.4. This highlights its capability to accurately estimate poses under moderate occlusions. Although the model yields a slightly lower AP_{hard} of 65.9 compared to CCAM-Person at 66.9 and MAQT at 66.7, it remains highly competitive. These results emphasize the contribution of adaptive multi-scale feature fusion and dual-gate context blocks in resolving complex occlusions and effectively separating overlapping human joints.

Our strong performance on CrowdPose's challenging occlusion scenarios validates AMFACPose's approach to occlusion handling through feature enhancement rather than explicit visibility classification as in HPCVNet [26]. While HPCVNet achieves slightly higher mAP on COCO (77.6 vs. 76.6), its performance on occlusion-heavy datasets like CrowdPose has not been established. The combination of multi-scale feature fusion via the AFPN and contextual refinement through DGCBs in our approach proves particularly effective for distinguishing overlapping subjects in real-world scenarios.

Altogether, AMFACPose combines high accuracy with low parameter complexity and reliable performance in crowded visual scenes. The integration of attention-guided multi-scale processing and a coordinate classification framework makes it particularly well suited for real-time applications and deployment in embedded environments, where both precision and efficiency are essential.

Table 3. Performance evaluation of SOTA 2D HPE models on CrowdPose; AP scores across easy, medium, and hard tiers.

Model	Backbone	AP	AP_{50}	AP_{75}	AP_{easy}	AP_{medium}	AP_{hard}
Sim.Baseline [7]	ResNet-50	60.8	81.4	65.7	71.4	61.2	51.2
HRNet [8]	HigherHRNet-W48	67.6	87.4	72.6	75.8	68.1	58.9
DEKR [51]	DEKR-W48	68.0	85.5	73.4	76.6	68.8	58.4
ED-Pose [52]	ResNet-50	69.9	88.6	75.8	77.7	70.6	60.9
PRTR [50]	Swin-L	71.6	90.4	78.3	77.3	72.0	65.8
HRFormer [9]	HRFormer-B	72.6	89.6	77.2	76.6	73.5	59.5
ED-Pose [52]	Swin-L	73.1	90.5	79.8	80.5	73.8	63.8
HDA-Pose [53]	YOLOv8	73.7	92.3	77.5	79.8	75.0	66.1
GroupPose [54]	Swin-L	74.1	91.3	80.4	80.8	74.7	66.4
MAQT [55]	HRNet-S	74.3	91.5	80.5	80.9	74.8	66.7
CCAM-Person [56]	YOLOv8	74.4	92.7	78.4	80.4	75.7	66.9
AMFACPose	**ResNet34**	**75.3**	**93.4**	**81.0**	**82.1**	**76.4**	65.9

Bold values indicate the proposed method and the best performance in each evaluation metric.

5.4. Qualitative Analysis

Figure 8 depicts representative AMFACPose outputs on the COCO dataset, demonstrating the framework's adaptability to diverse poses, occlusions, and environmental conditions. In high-action sports scenarios such as tennis, the model accurately tracks rapid limb movements without sacrificing fine-grained joint precision. Outdoor settings, ranging from skateboarding parks to beach environments, highlight the system's resilience to shifting backgrounds, lighting variations, and dynamic body configurations. Even in multi-person scenes where individuals overlap, AMFACPose reliably distinguishes each subject's joints, showing the effectiveness of its multi-scale fusion and gating modules. These qualitative observations align with the quantitative metrics reported earlier, suggesting AMFACPose's potential for real-world deployment in applications demanding both accuracy and computational efficiency.

Figure 8. Qualitative examples of AMFACPose's predictions on COCO images under diverse scenarios, including fast sports movements, overlapping subjects, and challenging lighting conditions. Results illustrate model's robustness to occlusions, pose variations, and multi-person interactions.

To further investigate the effectiveness of our proposed DGCB module in handling occlusion, we visualize attention maps with and without the DGCB enabled. As shown in Figure 9, we evaluate a sample image where a subject's hands and shoulders are partially occluded by flowers. The attention maps indicate that when the DGCB is active, the model maintains more focused attention on the occluded joints. The difference map (rightmost panel) highlights the regions where attention increases (red) or decreases (blue), showing that the DGCB effectively enhances attention near occluded keypoints.

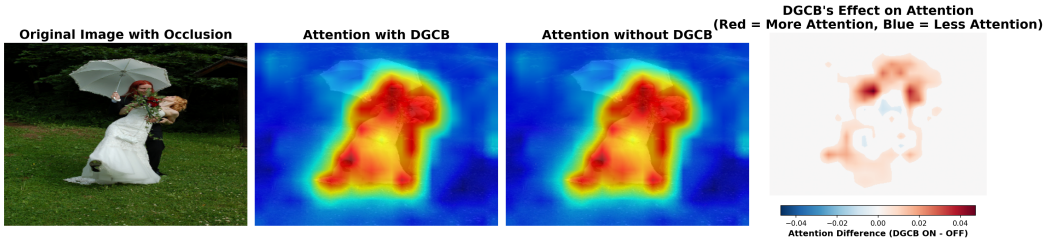

Figure 9. Visualization of DGCB's effectiveness in handling occlusion. (**Left**): Original occluded image. (**Center**): Attention maps with and without DGCB. (**Right**): Difference map (DGCB ON–OFF), showing improved focus in red and decreased attention in blue.

Figure 10 provides a detailed view of the occluded region. Zoomed attention maps clearly show that the DGCB helps maintain activation on partially hidden body parts. The additional focus around occluded limbs demonstrates how the DGCB leverages global context to refine local attention, improving keypoint localization under challenging visibility conditions.

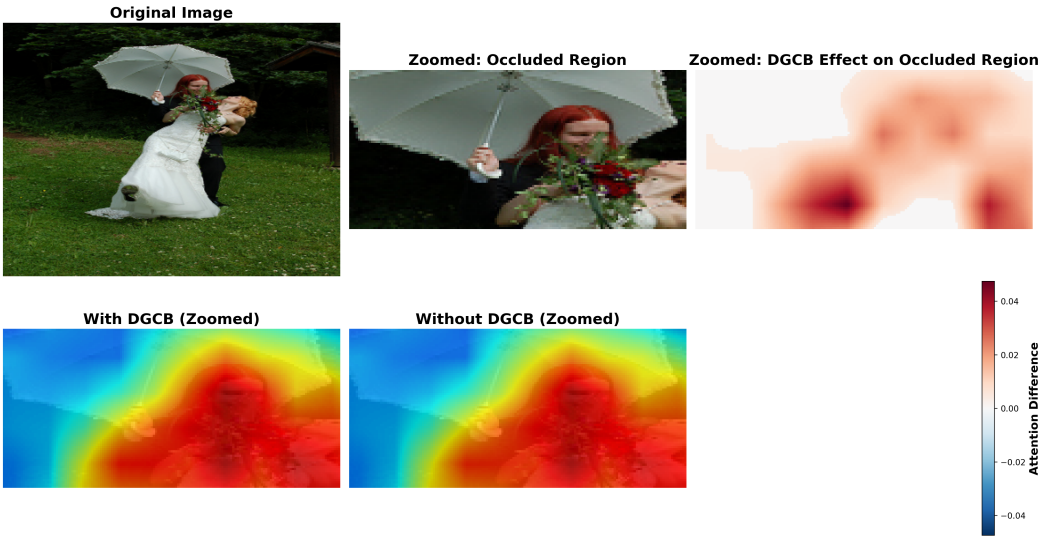

Figure 10. Zoomed qualitative analysis of DGCB's effect on occluded body parts. Top row: Original image and zoomed region. Bottom row: Attention with and without DGCB, showing stronger activation on occluded subject's upper body when DGCB is used.

These qualitative results reinforce our quantitative findings on the CrowdPose benchmark and provide visual evidence that DGCBs significantly enhance occlusion robustness in AMFACPose.

6. Performance and Efficiency Analysis

Figure 11 offers a visual overview of the speed–accuracy trade-offs for 2D HPE models, highlighting how AP correlates with inference speed on an RTX 3060 GPU. Each marker's size denotes the model's overall performance, while the green-shaded area represents the "sweet spot", where a favorable balance of high AP and real-time throughput emerges. AMFACPose, with a ResNet34 backbone, resides within this region, emphasizing its ability to maintain robust accuracy while achieving competitive frame rates. In comparison, alternative approaches often sacrifice accuracy for speed, or vice versa, reinforcing AMFACPose's balanced design.

Table 4 summarizes AMFACPose's performance across multiple hardware platforms, including the Jetson Orin Nano-8 at 15 W, Orin NX-8 at 20 W, Orin NX-16 at 25 W, and two desktop GPUs, which are RTX 4090 and RTX 3060. Using a ResNet34 backbone with a 384 × 288 input on the Orin Nano-8, this configuration processes each frame in 67.82 ms, corresponding to 14.74 fps. The more powerful Orin NX-16 reduces inference time to 45.97 ms and yields 21.75 fps. High-end GPUs offer even higher throughput, where the same setup runs at 6.74 ms per frame, i.e., 148.45 fps, on the RTX 4090 and 9.90 ms, i.e., 101.04 fps, on the RTX 3060. Reducing the input resolution to 256 × 256 substantially increases speed, especially on lower-wattage devices. For instance, the ResNet34 variant at 256 × 256 runs at 23.59 fps on the Orin Nano-8, scaling to over 30 fps on the Orin NX-8 and exceeding 100 fps on the RTX 3060.

Switching to a ResNet18 backbone trades off some precision for even greater efficiency. At a 384 × 288 input, this lighter configuration processes frames in 43.15 ms with 23.18 fps on the Orin Nano-8. This improves to 29.64 ms with 33.74 fps on the Orin NX-16, and further reduces to under 5 ms per frame with 201.10 fps on the RTX 4090. The 256 × 256 variant pushes the fps further across all devices; it reaches 36.02 fps on the Orin Nano-8 and scales to 50.45 fps on the Orin NX-16, while desktop GPUs offer well over 150 fps with modest sacrifices in accuracy relative to the ResNet34 backbone. These observations confirm that AMFACPose can flexibly adapt to various computational budgets, ranging

from highly constrained edge platforms to powerful desktop workstations, while retaining a competitive balance between speed and precision.

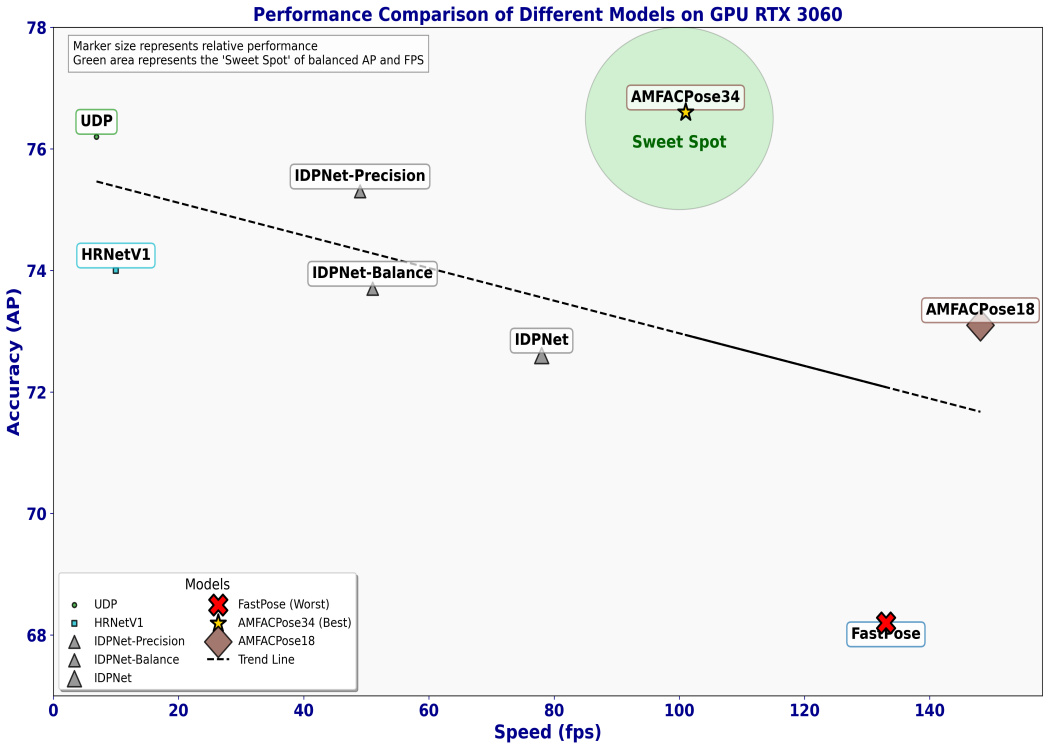

Figure 11. Speed–accuracy trade-offs for various models on RTX 3060 GPU. Each marker's size indicates relative performance, and green zone denotes "sweet spot" balancing high AP with fps. AMFACPose achieves notable accuracy in this region without incurring heavy computational costs.

Table 4. Performance analysis of AMFACPose model across different hardware platforms.

Model	Backbone	Input	AP	Orin Nano-8 (15 W) ms/fps	Orin NX-8 (20 W) ms/fps	Orin NX-16 (25 W) ms/fps	RTX 4090 ms/fps	RTX 3060 ms/fps
AMFACPose	ResNet34	384 × 288	76.6	67.82/14.74	52.90/18.90	45.97/21.75	6.74/148.45	9.90/101.04
		256 × 256	75.6	42.46/23.59	33.18/30.23	29.28/34.16	6.46/154.80	8.75/114.31
	ResNet18	384 × 288	73.1	43.15/23.18	33.05/29.89	29.64/33.74	4.97/201.10	6.75/148.19
		256 × 256	72.1	27.76/36.02	22.21/45.03	19.82/50.45	4.85/206.39	6.15/162.53

Table 5 places AMFACPose alongside prominent 2D HPE methods on an RTX 3060. Even with the ResNet34 backbone, the model obtains 101.0 fps, surpassing the speed of most competing solutions and matching or outperforming many in accuracy. The ResNet18 setup, although slightly lower in AP at 73.1, pushes throughput to 148.1 fps. Models like UDP with ResNet152 reach comparable precision but operate at only 6.9 fps, highlighting AMFACPose's efficiency advantages. The results of Table 5 are visualized in Figure 11. Across diverse settings and hardware, these results demonstrate that our architecture consistently maintains a favorable trade-off between accuracy and real-time usability, making it a versatile choice for applications that demand both precision and scalability.

Table 5. Comparison of AMFACPose with other methods on RTX_3060.

Model	Backbone	AP	GPU–RTX-3060 (fps)
UDP [28]	ResNet152	76.2	6.9
HRNetV1 [8]	HRNetV1-w32	74.0	10.0
IDPNet-Precision [20]	IDPNet-1286	75.3	49.0
IDPNet-Balance [20]	IDPNet-1143	73.7	51.0
IDPNet [20]	DPNet-1132	72.6	78.0
FastPose [19]	LPN-50	68.2	133.0
AMFACPose	ResNet34	76.6	101.0
	ResNet18	73.1	148.1

7. Ablation Study

A systematic ablation study was conducted to evaluate the contribution of each core component in AMFACPose. Table 6 reports the results on the MS COCO dataset using a ResNet34 backbone with an input resolution of 384 × 288. Starting from the baseline, we incrementally added the AFPN, DGCBs, SCRM, SE blocks, and CoordConv2d layers, enabling a detailed assessment of both the accuracy and computational cost associated with each module.

Table 6. Ablation study analysis of performance and computational cost of AMFACPose components on COCO dataset.

Model Variant	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	Params (M)	FLOPs (G)
Baseline (ResNet34)	72.4	91.3	78.3	69.3	77.3	75.3	3.3	4.7
+AFPN	73.2	91.5	79.3	69.9	78.5	76.3	3.5	5.0
+DGCBs	73.4	91.5	79.9	70.1	78.8	76.9	3.6	5.0
+SCRM	75.4	92.5	82.0	72.6	80.2	78.8	3.6	5.2
+SE	76.3	92.6	82.8	73.0	81.2	79.0	3.8	5.2
+CoordConv2d **(Full AMFACPose)**	**76.6**	**92.6**	**83.7**	**73.9**	**81.2**	**79.3**	**3.8**	**5.2**

The bold is used to highlight our complete proposed method (Full AMFACPose).

The baseline configuration achieves an AP of 72.4 with 3.3 M parameters and 4.70 GFLOPs, serving as the initial reference. Adding the AFPN increases AP to 73.2 while slightly increasing the model size to 3.5 M parameters and 5.0 GFLOPs, demonstrating the benefit of efficient multi-scale feature aggregation. Introducing DGCBs further improves AP to 73.4, with parameters at 3.6 M and FLOPs at 5.0, highlighting better contextual encoding with minimal cost.

A more substantial performance gain comes from integrating the SCRM module, which increases AP to 75.4 while maintaining 3.6 M parameters and 5.2 GFLOPs. This shows the effectiveness of joint spatial and channel refinement. The inclusion of SE blocks lifts AP to 76.3, with a modest increase to 3.8 M parameters and no additional GFLOP cost, due to lightweight global re-weighting.

Finally, adding CoordConv2d brings the model to its full configuration, achieving an AP of 76.6, along with 92.6 AP_{50}, 83.7 AP_{75}, 73.9 AP_M, and 81.2 AP_L. The model remains compact with 3.8 M parameters and 5.2 GFLOPs.

Importantly, this analysis directly addresses concerns about the integration of multiple attention mechanisms. Across the entire architecture, AMFACPose improves AP by 4.2 while increasing parameter count by just 0.5 M and computation by 0.5 GFLOPs. These results confirm that the proposed modules—the AFPN, DGCBs, SCRM, and SE blocks—work collaboratively and efficiently, making the full model highly suitable for real-time and resource-constrained environments.

8. Conclusions and Future Work

This paper presents AMFACPose, an efficient architecture for 2D HPE that addresses key challenges in scale variation and occlusion while maintaining minimal computational requirements. Through the integration of specialized components—the AFPN, DGCBs, SE blocks, and SCRM—our coordinate-based classification approach achieves high localization accuracy without the computational overhead typical of heatmap-based methods. Extensive evaluations on the COCO and CrowdPose datasets demonstrate AMFACPose's effectiveness, outperforming state-of-the-art approaches while maintaining significantly lower parameter counts and faster inference speeds across various hardware platforms. The architecture's adaptability is particularly evident in its consistent performance across resource-constrained edge devices and high-performance GPUs.

Despite its advantages, AMFACPose—like most vision models—may experience performance degradation under extreme conditions such as low-light environments or severe motion blur. Addressing these limitations requires future extensions that integrate spatiotemporal information or leverage cross-modality learning. We also acknowledge the ethical implications of pose estimation technologies, particularly in surveillance contexts, where privacy concerns are critical. Responsible deployment must be guided by transparent governance, informed consent, and equitable data representation to avoid bias and ensure fairness across demographic groups.

Future research directions include extending AMFACPose to 3D pose estimation and multi-view configurations, integrating temporal information for dynamic motion analysis, and developing domain adaptation techniques for limited-data scenarios. Additional focus will be placed on power-optimization strategies for embedded and mobile deployments, further enhancing the model's practical utility in resource-constrained environments.

Author Contributions: Conceptualization, A.Z., S.A.S. and H.T.; Methodology, A.Z.; Software, A.Z. and S.A.S.; Validation, S.A.S. and G.B.-G.; Formal analysis, G.B.-G.; Investigation, A.Z.; Writing—review & editing, G.B.-G. and H.T.; Supervision, H.T.; Project administration, H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in this article.

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

Abbreviations

HPE	Human Pose Estimation
DL	Deep Learning
DSC	Depthwise Separable Convolutions
MLPs	Multi-Layer Perceptron
GAP	Global Average Pooling
AFPN	Adaptive Feature Pyramid Network
DGCB	Dual-Gate Context Block
SE	Squeeze-and-Excitation
SCRM	Spatial–Channel Refinement Module
CoordConv2d	Coordinate Convolution 2D
KLDiscretLoss	Kullback–Leibler Divergence-based Discrete Loss
MSE	Mean Squared Error
OKS	Object Keypoint Similarity
AP	Average Precision
AR	Average Recall
IoU	Intersection over Union

CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
GFLOPs	Giga FLoating-point Operation
fps	Frames Per Second
BN	Batch Normalization
ReLU	Rectified Linear Unit
COCO	Common Objects in Context
FPN	Feature Pyramid Network
SOTA	State-Of-The-Art

References

1. Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; Torresani, L. Learning Temporal Pose Estimation from Sparsely-Labeled Videos. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 3003–3014.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. Available online: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (accessed on 15 April 2025). [CrossRef]
3. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299. [CrossRef]
4. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660. [CrossRef]
5. Yang, Y.; Ramanan, D. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2878–2890. [CrossRef] [PubMed]
6. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Amsterdam, The Netherlands, 2016; pp. 483–499. [CrossRef]
7. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481. [CrossRef]
8. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703. [CrossRef]
9. 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]
10. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.-T.; Zhou, E. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 11313–11322. [CrossRef]
11. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/fbb10d319d44f8c3b4720873e4177c65-Paper-Conference.pdf (accessed on 10 April 2025).
12. Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; Xia, S.T. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 89–106. [CrossRef]
13. Zakir, A.; Salman, S.A.; Benitez-Garcia, G.; Takahashi, H. AECA-PRNetCC: Adaptive Efficient Channel Attention-Based PoseResNet for Coordinate Classification in 2D Human Pose. In Proceedings of the 38th International Conference on Image and Vision Computing New Zealand (IVCNZ), Auckland, New Zealand, 29 November–1 December 2023; pp. 1–6. [CrossRef]
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
15. Liu, R.; Lehman, J.; Molino, P.; Such, F.P.; Frank, E.; Sergeev, A.; Yosinski, J. An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018; pp. 9605–9616. [CrossRef]
16. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [CrossRef]

17. Li, J.; Bian, S.; Zeng, A.; Wang, C.; Pang, B.; Liu, W.; Lu, C. Human Pose Regression with Residual Log-Likelihood Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021; pp. 11025–11034. [CrossRef]
18. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1799–1807. [CrossRef]
19. Dai, H.; Shi, H.; Liu, W.; Wang, L.; Liu, Y.; Mei, T. FasterPose: A Faster Simple Baseline for Human Pose Estimation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *18*, 1–16. [CrossRef]
20. Liu, H.; Wu, J.; He, R. IDPNet: A Light-Weight Network and Its Variants for Human Pose Estimation. *J. Supercomput.* **2024**, *80*, 6169–6191. [CrossRef]
21. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-Aware Coordinate Representation for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 14–19 June 2020; pp. 7093–7102. [CrossRef]
22. Yin, S.; Wang, S.; Chen, X.; Chen, E.; Liang, C. Attentive One-Dimensional Heatmap Regression for Facial Landmark Detection and Tracking. In Proceedings of the 28th ACM International Conference on Multimedia (ACM MM), Seattle, WA, USA, 12–16 October 2020; pp. 538–546. [CrossRef]
23. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 14–19 June 2020; pp. 5700–5709. [CrossRef]
24. Zakir, A.; Salman, S.A.; Takahashi, H. SOCA-PRNet: Spatially Oriented Attention-Infused Structured-Feature-Enabled PoseResNet for 2D Human Pose Estimation. *Sensors* **2023**, *23*, 110. [CrossRef] [PubMed]
25. Zakir, A.; Salman, S.A.; Benitez-Garcia, G.; Takahashi, H. EBA-PRNetCC: An Efficient Bridge Attention-Integration PoseResNet for Coordinate Classification in 2D Human Pose Estimation. In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Rome, Italy, 27–29 February 2024; pp. 133–144. [CrossRef].
26. Jiang, Z.; Ji, H.; Yang, C.-Y.; Hwang, J.-N. 2D Human Pose Estimation Calibration and Keypoint Visibility Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 6095–6099. [CrossRef]
27. 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 European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
30. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; pp. 87.1–87.12. [CrossRef]
31. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
32. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980. [CrossRef]
33. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
34. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human Pose Estimation with Iterative Error Feedback. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4733–4742. [CrossRef]
35. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
36. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
37. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 1321–1330. [CrossRef]

38. PyTorch Documentation. *torch.nn.KLDivLoss*. Available online: https://pytorch.org/docs/stable/generated/torch.nn.KLDivLoss.html (accessed on 1 February 2025).
39. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10863–10872. [CrossRef]
40. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
41. Misra, D. Mish: A Self Regularized Non-Monotonic Activation Function. *arXiv* **2019**, arXiv:1908.08681. [CrossRef]
42. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for Activation Functions. *arXiv* **2017**, arXiv:1710.05941. [CrossRef]
43. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2017. [CrossRef]
44. Zhang, M.; Lucas, J.; Ba, J.; Hinton, G.E. Lookahead Optimizer: k Steps Forward, 1 Step Back. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 9597–9608. [CrossRef]
45. Bello, I.; Zoph, B.; Vasudevan, V.; Le, Q.V. Neural Optimizer Search with Reinforcement Learning. In Proceedings of the International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 459–468. [CrossRef]
46. Liu, Z.; Liu, L.; Hao, S. BR-Pose: Enhancing Human Pose Estimation Through Bi-level Routing Attention and Multi-level Weight Fusion. *Vis. Comput.* **2025**, 1–12. [CrossRef]
47. Tian, Z.; Fu, W.; Woźniak, M.; Liu, S. PCDPose: Enhancing the Lightweight 2D Human Pose Estimation Model with Pose-enhancing Attention and Context Broadcasting. *Pattern Anal. Appl.* **2025**, *28*, 59. [CrossRef]
48. Chen, S.; Zhang, Y.; Huang, S.; Yi, R.; Fan, K.; Zhang, R.; Chen, P.; Wang, J.; Ding, S.; Ma, L. SDPose: Tokenized Pose Estimation via Circulation-Guide Self-Distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 1082–1090. [CrossRef]
49. Zhang, F.; Shi, Q.; Ma, Y. Combining Self-attention and Depth-wise Convolution for Human Pose Estimation. *SIViP* **2024**, *18*, 5647–5661. [CrossRef]
50. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose Recognition with Cascade Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1944–1953. [CrossRef]
51. Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; Wang, J. Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14676–14686. [CrossRef]
52. 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. [CrossRef]
53. Dong, C.; Tang, Y.; Zhang, L. HDA-Pose: A Real-Time 2D Human Pose Estimation Method Based on Modified YOLOv8. *Signal Image Video Process.* **2024**, *18*, 5823–5839. [CrossRef]
54. Liu, H.; Chen, Q.; Tan, Z.; Liu, J.-J.; Wang, J.; Su, X.; Li, X.; Yao, K.; Han, J.; Ding, E.; et al. GroupPose: A Simple Baseline for End-to-End Multi-Person Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 15029–15038. [CrossRef]
55. Liang, H.; Wang, C.; Shao, M.; Zhang, Q. MAQT: Multi-scale attention and query-optimized transformer for end-to-end pose estimation. *J. Supercomput.* **2025**, *81*, 429. [CrossRef]
56. Dong, C.; Du, G. An enhanced real-time human pose estimation method based on modified YOLOv8 framework. *Sci. Rep.* **2024**, *14*, 8012. [CrossRef] [PubMed]