Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(9): 5861–5886.
DOI: 10.3934/era.2026260
Received: 25 April 2026
Revised: 26 June 2026
Accepted: 30 June 2026
Published: 09 July 2026

Research article

FDC-pose: Enhancing multi-person pose estimation in complex scenes via dynamic synergy mechanisms

Mengye Lin[1], Zhiming Cai[1,2,*], Yixin Zhang[1], Jiangchao Zhang[1], Zukun Xu[1] and Huabin He[1,2]

[1] School of Electrical and Information Engineering, Fujian University of Technology, Fuzhou 350118, China

[2] National Demonstration Center for Experimental Electronic Information and Electrical Technology Education, Fujian University of Technology, Fuzhou 350118, China

* **Correspondence:** Email: caizm@fjut.edu.cn.

Abstract: Real-time multi-person pose estimation suffers from inter-instance occlusion and severe scale variations. To resolve these bottlenecks, we present focal-dynamic context (FDC)-pose, an efficient single-stage architecture tailored for crowded scenes. The framework incorporates three sequential innovations: 1) a cascaded pyramid network-focal omni-dimensional dynamic convolution (CPN-FocalOD) backbone combining context modulation with omni-dimensional dynamic filtering to disentangle overlapping silhouettes; 2) a cross-scale fusion network featuring content-aware upsampling (DySample) to suppress quantization errors and preserve distal joint details; and 3) a minimum point distance intersection over union (MPDIoU) regression loss enforcing point-level geometric constraints to eliminate occlusion-induced localization jitter. Evaluations on the common objects in context (COCO) 2017 dataset demonstrate that FDC-pose yields absolute gains of 3.4% in AP, 3.0% in AP_{50}, and 4.2% in the strict AP_{75} metric relative to the baseline. Parallel evaluations on the CrowdPose dataset show that our framework presents a prominent robustness gain, yielding an absolute improvement of 4.5% in AP.

Keywords: human pose estimation; YOLO; dynamic convolution; multi-scale feature fusion; content-aware upsampling; bounding box regression

1. Introduction

Driven by advances in convolutional neural networks (CNNs) [1] and object detection frameworks [2], monocular human pose estimation (HPE) has become a fundamental computer vision task aimed at localizing specific anatomical keypoints, such as shoulders, elbows, and knees, from 2D images [3]. As a technological cornerstone, HPE supports an extensive range of downstream

applications, encompassing traditional domains like action recognition [4] and motion evaluation. In affective computing, specifically, spatiotemporal pose dynamics provide indispensable body language cues; when fused with facial and physiological signals, these cues significantly enhance the robustness of multimodal emotion recognition systems. Furthermore, providing high-fidelity 2D geometric joint configurations directly benefits higher-dimensional tasks, establishing a vital prerequisite for downstream monocular 3D pose reconstruction via interactive complementary feature fusion [5].

The evolution of HPE architectures reflects a continuous trade-off between spatial precision and computational overhead. Early milestones, such as stacked hourglass [6] and simple baselines [7], established robust spatial modeling via cascaded modules and deconvolutional layers. Subsequent innovations like the cascaded pyramid network (CPN) [8] and high-resolution network (HRNet) [9] further refined localization by handling invisible keypoints and maintaining parallel high-resolution representations. Recently, vision transformers, exemplified by ViTPose [10], have introduced global attention mechanisms to the field; meanwhile, advanced dual-domain cross enhanced transformer topologies have also been successfully applied to handle complex topological couplings in higher-dimensional spatial spaces [11]. However, their prohibitive computational costs restrict deployment in resource-constrained real-time applications [12].

Beyond individual architectures, multi-person HPE paradigms predominantly follow either top-down [13] or bottom-up [14] strategies. Top-down methods rely on a cascaded "detect-then-estimate" process, where computational latency scales linearly with the number of instances, creating severe bottlenecks in crowded scenes. Conversely, bottom-up approaches such as OpenPose [15] struggle with NP-hard grouping processes, often leading to association errors under severe occlusion. To circumvent these limitations, you only look once (YOLO)-based single-stage regression frameworks have emerged, redefining pose estimation to establish a robust foundation for real-time applications [16].

Despite their inference efficiency, single-stage detectors often compromise on feature granularity and multi-scale fusion, exhibiting degraded robustness in uncontrolled environments characterized by severe occlusion or motion blur [17]. Through a theoretical analysis of the baseline framework, we identify three core structural bottlenecks constraining model performance in challenging scenarios. First, standard convolutions apply fixed kernel weights indiscriminately. This non-adaptive nature lacks the content awareness required to decouple overlapping physical patterns and handle non-rigid deformations or severe occlusions, thereby weakening local spatial precision [18, 19]. Second, standard feature pyramid networks (FPNs) typically rely on static upsampling operators such as nearest-neighbor interpolation. These content-agnostic operators induce severe edge smoothing and information loss during scale recovery. Furthermore, semantic and geometric misalignments across feature levels constrain the accurate reconstruction of minute distal joints [20, 21]. Finally, existing loss functions like complete intersection over union (CIoU) [22] rely on global shape attributes, including center distance and aspect ratio, but lack strict pixel-level geometric constraints. In occlusion-prone environments, this deficiency causes "localization jitter", where bounding box misalignments propagate spatial errors directly to the keypoint regression head.

To surmount these architectural limitations, we propose the focal-dynamic context pose estimation network, termed FDC-pose, which is an efficient single-stage framework explicitly optimized for complex spatial contexts. Specifically, the network resolves the static feature extraction bottleneck by constructing the cascaded pyramid network-focal omni-dimensional dynamic convolution

(CPN-FocalOD) backbone. By synergizing global context modulation with omni-dimensional dynamic filtering, this module spatially calibrates kernel weights based on instance-specific variations, fundamentally enhancing the network's ability to disentangle complex, overlapping anatomical features in crowded scenes. To mitigate upsampling degradation in the neck stage, we introduce a dual synergistic strategy where the dynamic sampling (DySample) content-aware point sampling operator replaces static interpolation to restore high-resolution representations without blurring fine-grained details. Concurrently, the cross-scale feature fusion module (CCFM) employs a residual-based dual-path architecture to strictly align and aggregate multi-scale semantics and shallow geometric cues, effectively restoring localization precision for small-scale joints. Finally, geometric localization ambiguity is addressed by integrating the minimum point distance intersection over union (MPDIoU) as the core regression loss. By directly minimizing the Euclidean distance between the corner points of predicted and ground-truth boxes, MPDIoU imposes stringent geometric constraints, which effectively penalizes aspect ratio distortion and eliminates occlusion-induced localization jitter, establishing a precise geometric foundation that blocks error propagation to the keypoint head.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents the experiments and ablation studies. Section 5 discusses the findings, and Section 6 concludes the paper.

2. Related work

2.1. Human pose estimation paradigms

HPE paradigms are generally categorized into top-down, bottom-up, and single-stage approaches, each balancing localization precision and computational efficiency differently [23]. Top-down methods normalize targets into local high-resolution crops, ensuring fine-grained anatomical capture. However, their inference latency scales linearly with the number of persons, prohibiting real-time deployment on edge devices. Conversely, bottom-up paradigms employ a parallel "detect-then-group" mechanism. While efficient in multi-person scenarios, they struggle with NP-hard grouping processes under severe occlusion, often leading to irreversible matching errors. To combine the strengths of both, single-stage regression paradigms, particularly YOLO-based architectures [24], unify detection and estimation into an end-to-end mapping, offering a superior speed-accuracy trade-off. Nevertheless, standard YOLO backbones rely on static convolutions with fixed kernel weights. This content-agnostic paradigm lacks the contextual adaptability required to disentangle overlapping limbs in crowded scenes [25]. To address this feature extraction bottleneck, we propose the CPN-FocalOD module. By integrating Focal Modulation [26] and omni-dimensional dynamic convolution (ODConv) [19], it replaces static convolution units to provide omni-dimensional dynamic context perception, thereby enhancing instance-specific feature reconstruction.

2.2. Multi-scale feature aggregation and upsampling

Handling scale variance necessitates effective multi-scale feature aggregation. Although modern detectors utilize path aggregation networks (PANets) [27], they still suffer from dual structural bottlenecks when processing fine-grained joints. First, traditional FPNs rely heavily on static, content-agnostic upsampling operators such as nearest-neighbor interpolation, which induce feature

blurring and distort minute distal joints during resolution recovery [28]. Second, simple cascade fusion mechanisms fail to align significant semantic and geometric discrepancies across levels, limiting discrimination accuracy under complex poses. To mitigate these degradation issues, our Neck stage employs a dual synergistic strategy. We integrate DySample [29], a content-aware dynamic point sampling mechanism, to strictly preserve fine-grained anatomical details during upsampling. Concurrently, the CCFM is introduced to adaptively align and aggregate multi-scale feature flows via a residual-based dual-path architecture. Together, they resolve semantic gaps and provide robust representations for drastic scale variations.

2.3. Geometric-consistent bounding box optimization

Robust pose estimation relies heavily on precise bounding box regression, as its localization accuracy defines the effective receptive field for subsequent feature extraction. Optimization metrics have evolved from L_n-norm losses to intersection over union (IoU) variants such as generalized intersection over union (GIoU) [30], distance intersection over union (DIoU), and CIoU. CIoU has become the mainstream baseline in YOLO architectures due to its balance of convergence speed and localization performance. However, existing IoU metrics based on global shape attributes lack strict geometric discrimination. When predicted and ground-truth boxes share similar overlap rates but distinct spatial arrangements, these metrics fail to penalize misalignments effectively [31]. In occluded environments, this deficiency causes "localization jitter," preventing the bounding box from tightly fitting the skeletal contour and propagating spatial errors directly to the keypoint regression head. To establish a stringent geometric constraint, we adopt MPDIoU [32] as the core regression loss. Unlike CIoU, MPDIoU directly minimizes the Euclidean distance between the top-left and bottom-right corners of the predicted and ground-truth boxes. This point-level explicit alignment penalizes spatial misalignment and aspect ratio distortion, fundamentally suppressing occlusion-induced localization jitter and blocking error propagation to the keypoint head.

3. Method

To address content perception deficiency, scale mismatch, and geometric localization ambiguity in HPE, we propose the FDC-pose framework. This architecture integrates dynamic context modeling, multi-scale feature aggregation, and geometric-consistent regression to balance detection accuracy and computational efficiency in occluded and highly dynamic scenes. The overall network topology is illustrated in Figure 1.

The FDC-pose framework comprises three core sequential stages. In the backbone, standard static convolutions are replaced with the CPN-FocalOD module. By utilizing ODConv, this module adaptively recalibrates feature weights across spatial, channel, and kernel dimensions, mitigating background noise and enhancing the representation of high-information skeletal topologies under self-occlusion. Following feature extraction, the neck architecture employs a dual optimization mechanism integrating the CCFM and the DySample operator.

To further elucidate the detailed data flow and architectural logic within the network topology illustrated in Figure 1, the input tensor fed into the convolutional layer of the top-most CCFM block within the neck module is precisely configured as a channel-wise concatenated representation. This integrated tensor directly aggregates the fine-grained lateral feature maps from the shallow stage of

the CPN-FocalOD backbone and the dynamically upsampled semantic feature flows generated by the preceding DySample operator, ensuring an accurate qualitative alignment before being processed by the initial 1×1 convolution.

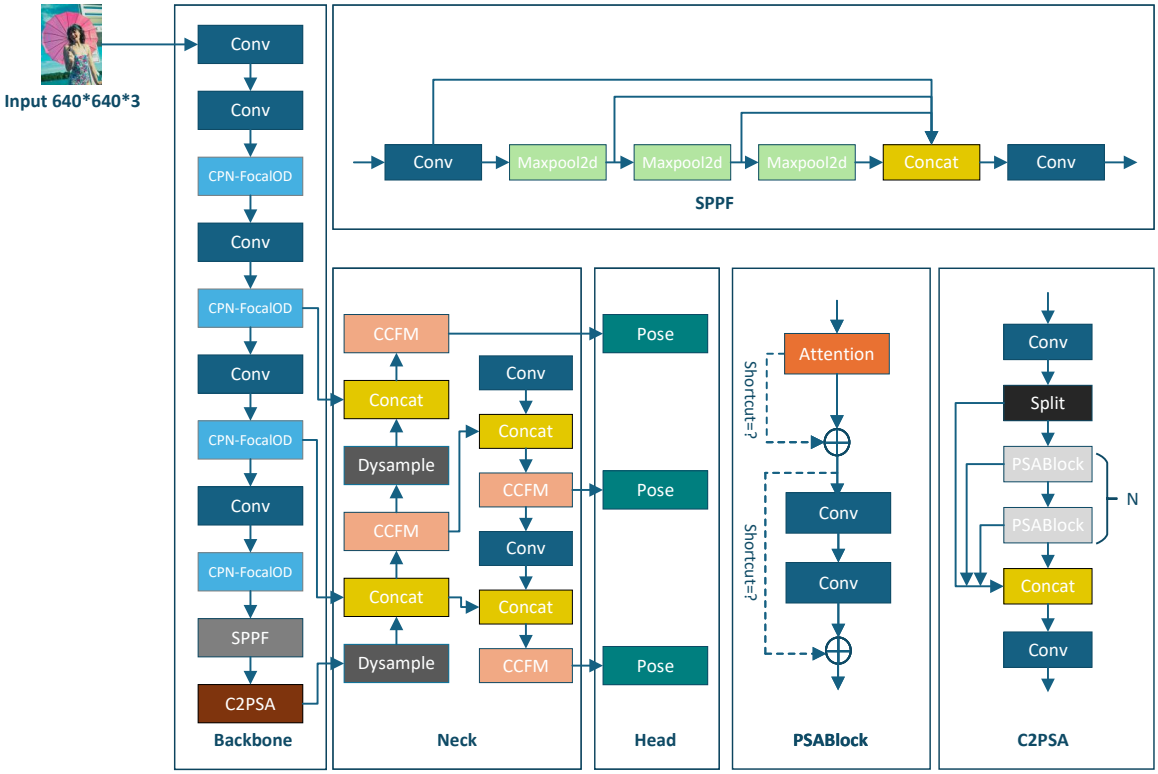


Figure 1. Network architecture of FDC-pose.

The CCFM bridges the semantic-geometric gap via a residual-based dual-path architecture, preserving small-scale joint structures. Concurrently, DySample replaces standard linear interpolation with content-aware point-wise offsets, rectifying quantization errors and preserving sharp structural boundaries during upsampling. In the prediction head, we adopt the MPDIoU loss to enforce explicit point-level geometric constraints. By directly minimizing the Euclidean distance between bounding box corners, this regression strategy eliminates localization jitter and blocks the cascading propagation of spatial errors to the keypoint regression head [33].

3.1. Baseline architectural configuration

To provide a comprehensive structural reference for our modifications, we delineate the hierarchical composition of the native YOLOv11-pose framework utilized as the baseline. The standard architecture operates on a unified single-stage regression paradigm that concurrently handles object detection and keypoint localization. Structurally, the backbone stage consists of standard static convolutional layers, standard C3k2 blocks for cross-stage partial feature extraction, a spatial pyramid pooling fast layer to capture multi-scale contextual fields, and a paired self-attention block designated as C2PSA to model long-range spatial dependencies. This assembly downsamples the input image

into hierarchical feature spaces, transferring abstract representations to the subsequent stages.

Following feature extraction, the native neck stage relies on a conventional PANet configuration. This topology orchestrates a symmetrical feature pyramid layout consisting of a sequential top-down pathway and a bottom-up pathway, which facilitates the bidirectional flow of high-level semantic data and shallow geometric localization cues. Within this neck network, spatial resolution recovery is executed using standard, content-agnostic nearest-neighbor interpolation layers. Ultimately, the aggregated multi-scale features are routed to decoupled prediction heads, where keypoint regression and bounding box tracking are optimized using standard object keypoint similarity (OKS) heads and the native CIoU loss function. This established baseline structure provides the fundamental feature representation and spatial tracking pathways upon which our custom algorithmic interventions are deployed.

3.2. CPN-FocalOD Module

Standard static convolutions struggle with the severe non-rigid deformations inherent in human poses. To extract high-fidelity representations under multi-person interference, the proposed CPN-FocalOD module employs a dual-branch architecture that decouples semantic context modeling from dynamic geometric adaptation. The detailed structure is shown in Figure 2.

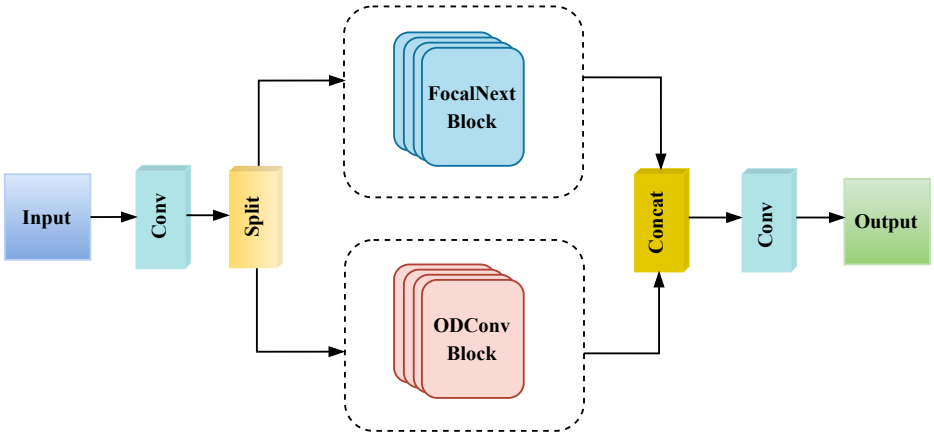


Figure 2. Detailed architecture of the CPN-FocalOD module.

Given an input tensor $X \in \mathbb{R}^{C\times H\times W}$, the module first processes it through an initial convolution layer. Subsequently, it splits the intermediate feature map along the channel dimension into a semantic subspace $X_{sem} \in \mathbb{R}^{\frac{C}{2}\times H\times W}$ and a dynamic subspace $X_{dyn} \in \mathbb{R}^{\frac{C}{2}\times H\times W}$. This equal division ratio is introduced to balance the computational resource allocation between context modeling and geometric filtering. This equal division ratio balances the feature capacity between global context learning and dynamic geometric adaptation. Specifically, allocating excessive channels to the semantic branch inevitably reduces the structural flexibility of the dynamic branch, which degrades the localization accuracy of small distal joints. Conversely, diminishing the semantic branch's capacity weakens the network's ability to distinguish overlapping human silhouettes in crowded environments. Thus, the 1:1 ratio serves as an optimal balance point to prevent performance degradation in either task. In multi-person pose scenes, the framework must concurrently separate overlapping physical silhouettes and localize fine-grained joints. Allocating an equivalent channel

width to each sub-pathway prevents the backbone from over-fitting to either task. This symmetric design effectively mitigates both the risk of feature undersampling in the dynamic branch under viewpoint changes and the risk of detail blurring in the semantic branch under background clutter, maintaining a robust feature representation across diverse scenes. This bifurcation allows independent specialization for distinct kinematic attributes, as processing multidimensional representations through decoupled evaluation pathways significantly suppresses latent noise interference [34].

The first branch processes X_{sem} via the FocalNext block (Figure 3), inspired by ConvNeXt [35]. It integrates dual skip connections with dilated depth-wise convolutions to capture both fine-grained local interactions and global context. Specifically, X_{sem} is processed by a 7×7 convolution and a DropPath [36] operation, followed by residual fusion to yield X_1. X_1 is further processed by a second 7×7 convolution, a DropPath operation, permutation, and layer normalization, followed by residual fusion to produce X_2. A projection layer comprising a 1×1 convolution and Gaussian error linear unit (GELU) activation maps X_2 to capture non-linear channel dependencies, yielding X_3. Finally, X_3 undergoes a terminal 1×1 convolution, a permutation operation, and a concluding DropPath. This output is then fused with the initial X_{sem} via a long-range residual connection, generating the final output X_{sem}^{out}. This hierarchical aggregation filters background noise while maintaining structural integrity, embodying a synergistic alignment that correlates opposing feature distributions to robustly preserve foundational topologies [37].

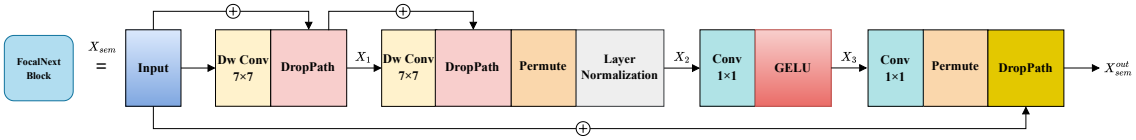


Figure 3. Detailed internal structure of the FocalNext block.

Concurrently, the second branch processes X_{dyn} using ODConv (Figure 4) to handle the geometric scale variations of human joints.

Structurally, X_{dyn} is compressed via global average pooling (GAP) to encode pose context, then mapped to a lower-dimensional space via a fully connected (FC) layer and rectified linear unit (ReLU) activation. Four parallel heads, each comprising an FC layer and a sigmoid activation, generate attention scalars to modulate the convolution along distinct dimensions: spatial dimensions (a_{si}), input channels (a_{ci}), output filters (a_{fi}), and kernel space (a_{wi}). The ODConv is mathematically formulated as:

$$X_{dyn}^{out} = \sum_{i=1}^{n} (W_i \odot a_{si} \odot a_{ci} \odot a_{fi} \odot a_{wi}) * X_{dyn}, \tag{3.1}$$

where X_{dyn}^{out} denotes the output tensor of the dynamic branch, W_i denotes the i-th convolution kernel, $\odot$ represents element-wise multiplication, and $*$ denotes the convolution operation. Synergizing these multidimensional attentions allows the dynamic kernel to adaptively adjust its receptive field. Spatial attention precisely localizes minute distal joints, while kernel attention manages limb scale fluctuations induced by viewpoint changes, effectively resolving complex occlusions.

To synthesize global semantic consistency with local geometric adaptability, the branch outputs are concatenated and fused via a mixing convolution:

$$F_{out} = \text{Conv}_{mix}(\text{Concat}(X_{sem}^{out}, X_{dyn}^{out})). \tag{3.2}$$

This fusion ensures F_{out} retains robust semantics while adapting to complex pose deformations, improving localization precision in challenging scenarios.

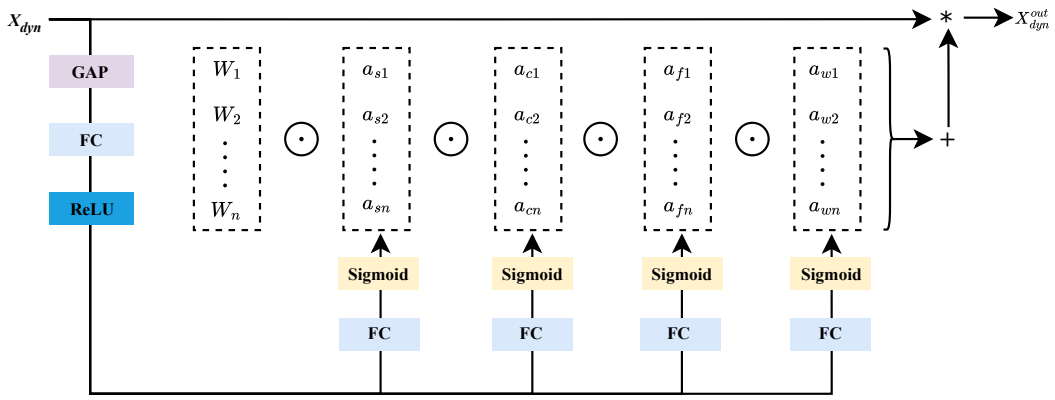


Figure 4. Architecture diagram of ODConv.

3.3. Cross-scale feature fusion and dynamic upsampling

High-precision HPE requires the robust aggregation of abstract semantic information with low-level geometric details to handle extreme scale variance, from macroscopic torso structures to microscopic distal joints. To bridge the semantic misalignment across different resolution stages, we introduce the CCFM. Inspired by the real-time detection transformer (RT-DETR) [38], the CCFM employs a residual-based dual-path architecture to achieve deep fusion of multi-scale features.

As depicted in Figure 5, rather than employing a simple cascading strategy,the CCFM first concatenates the input multi-scale features along the channel dimension. The concatenated feature map is then processed through a dual-branch topology. The first branch acts as a semantic shortcut, applying a 1×1 convolution for channel projection while preserving inherent feature identity. The second branch further refines the concatenated features by sequentially applying a 1×1 convolution and the computationally efficient C3k2 module, adapted for the YOLOv11 [39] framework. The outputs of both branches are aggregated via element-wise addition to generate the final fused representation. The complete fusion process is mathematically formulated as:

$$F_{fused} = \text{Conv}_{1\times1}(F_{cat}) + \text{C3k2}(\text{Conv}_{1\times1}(F_{cat})), \tag{3.3}$$

where F_{cat} denotes the initially concatenated multi-scale features. This residual aggregation ensures that robust semantic cues are injected into shallow layers while mitigating gradient degradation.

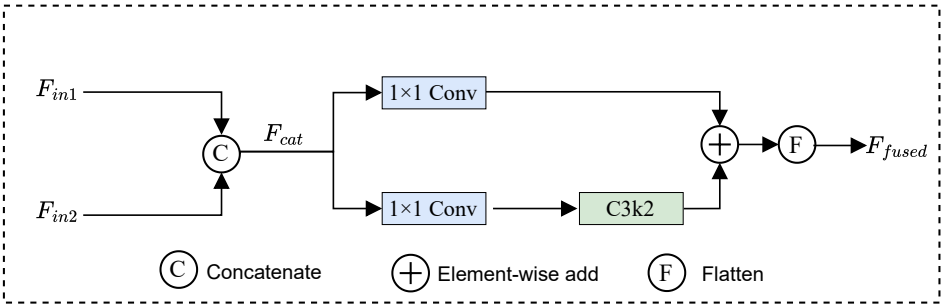
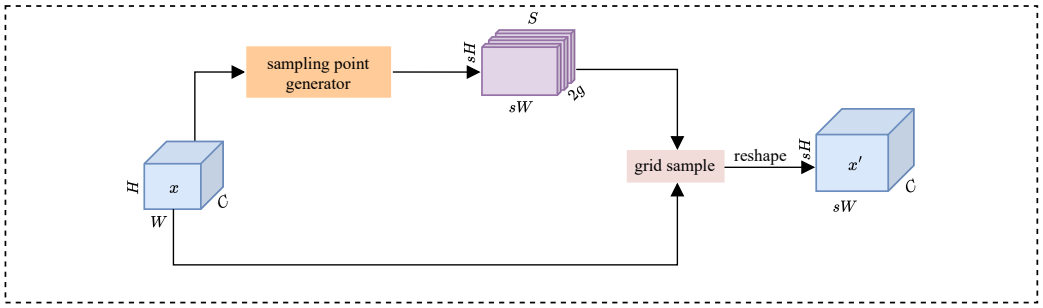
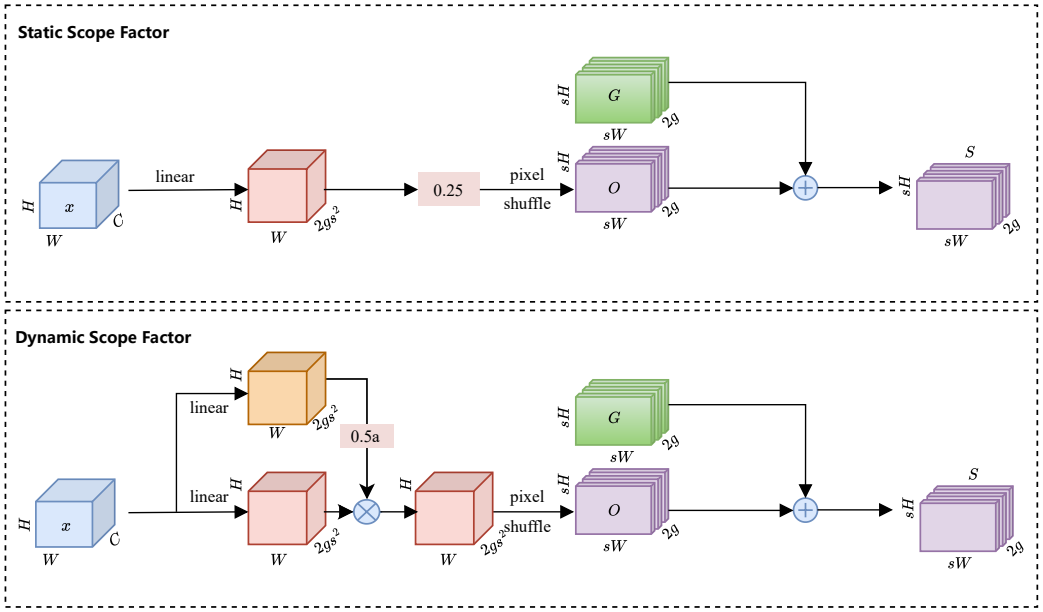


Figure 5. Detailed architecture of the CCFM.



(a) Sampling based dynamic upsampling



(b) Sample point generator in DySample

Figure 6. Schematic of the DySample module.

To reconstruct high-resolution feature maps without exacerbating computational overhead or blurring fine-grained distal joint details, we replace standard static interpolation with DySample (Figure 6), a content-aware dynamic upsampler. Given an input feature map $x \in \mathbb{R}^{C\times H\times W}$ and an upsampling scale factor s, a linear layer maps the channel dimension from C to $2gs^2$ (where g represents the number of groups). After being scaled by a scope factor, the PixelShuffle algorithm

reshapes the tensor into the offset map $O \in \mathbb{R}^{2g\times sH\times sW}$. These learned offsets are then superimposed onto the original sampling grid G to formulate the final sampling set S:

$$S = G + O. \tag{3.4}$$

Finally, the upsampled feature map $x' \in \mathbb{R}^{C\times sH\times sW}$ is generated via bilinear interpolation using the `grid_sample` function:

$$x' = \text{grid\_sample}(x, S). \tag{3.5}$$

This point-based dynamic sampling mechanism strictly preserves sharp structural boundaries during resolution recovery, optimizing the trade-off between model efficiency and localization precision.

3.4. MPDIoU

Precise bounding box regression is a prerequisite for single-stage keypoint regression, as it explicitly defines the effective receptive field and spatial anchor for the pose regression head [40]. While the standard CIoU incorporates center distance and aspect ratio constraints, it often fails to provide sufficient geometric discrimination under severe boundary ambiguity and occlusion. To enforce stricter alignment, we adopt MPDIoU as the bounding box regression objective. MPDIoU directly minimizes the normalized squared Euclidean distances between the corresponding corners (top-left and bottom-right) of the predicted and ground-truth boxes, explicitly penalizing spatial misalignment. The calculation flow is summarized in Algorithm 1.

Algorithm 1 Calculation flow of MPDIoU

Input: Two arbitrary convex shapes: $A, B \in S \subseteq \mathbb{R}^n$; Width and height of input image: w, h
Output: MPDIoU

1: For A and B, $(x_1^A, y_1^A), (x_2^A, y_2^A)$ denote the top-left and bottom-right point coordinates of A, respectively, while $(x_1^B, y_1^B), (x_2^B, y_2^B)$ denote the top-left and bottom-right point coordinates of B, respectively.
2: $d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2$
3: $d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2$
4: $MPDIoU = \frac{A\cap B}{A\cup B} - \frac{d_1^2}{w^2+h^2} - \frac{d_2^2}{w^2+h^2}$
5: **return** $MPDIoU$

Formally, let $B_{gt} = [x_1^{gt}, y_1^{gt}, x_2^{gt}, y_2^{gt}]$ and $B_{prd} = [x_1^{prd}, y_1^{prd}, x_2^{prd}, y_2^{prd}]$ denote the ground-truth and predicted bounding boxes, respectively. The squared Euclidean distances between their corresponding corners, denoted as d_1^2 and d_2^2, are calculated as:

$$d_1^2 = (x_1^{prd} - x_1^{gt})^2 + (y_1^{prd} - y_1^{gt})^2, \quad d_2^2 = (x_2^{prd} - x_2^{gt})^2 + (y_2^{prd} - y_2^{gt})^2. \tag{3.6}$$

Given the input image width w and height h, the MPDIoU metric incorporates a scale-invariant geometric penalty:

$$MPDIoU = IoU - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2}. \tag{3.7}$$

The final regression loss $\mathcal{L}_{MPDIoU}$ is defined as:

$$\mathcal{L}_{MPDIoU} = 1 - MPDIoU. \tag{3.8}$$

Minimizing Equation 3.8 forces explicit pixel-level alignment between the detection box and human contours. This significantly enhances robustness against occlusion and provides a precise geometric foundation for subsequent keypoint regression tasks.

4. Experiments and details

4.1. Experimental dataset

To comprehensively validate the performance and generalization capabilities of the proposed FDC-pose model, experiments were conducted on the common objects in context (COCO) 2017 keypoint dataset [41]. As a keypoint detection subset of the widely used COCO dataset, it serves as a standard and extensively recognized benchmark for modern HPE research.

This large-scale dataset comprises over 250,000 annotated human instances, each characterized by 17 anatomical keypoints. As illustrated in Figure 7, these keypoints define the skeletal topology ranging from facial landmarks to distal limb joints. For the experimental setup, we utilized the official data splits, which include 56,599 images for training, 2346 images for validation, and 20,000 images for testing. These subsets were employed for the initial training, evaluation, and final benchmarking of FDC-pose, respectively. The dataset is renowned for covering diverse scenarios involving both single and multi-person instances, drastic scale variations, and complex backgrounds, making it a critical platform for assessing model robustness under real-world conditions.

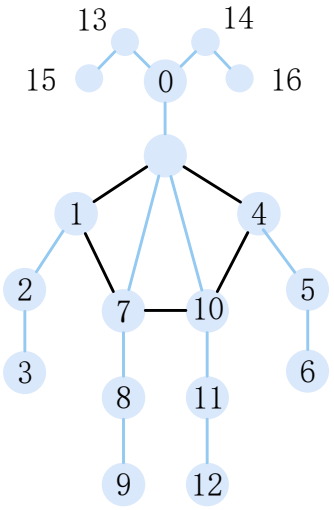


Figure 7. Standard 17-keypoint topology defined in the COCO dataset.

To further evaluate the environmental resilience and cross-dataset generalizability of FDC-pose under complex multi-person interactions, we extended our evaluation pipeline onto the authoritative CrowdPose dataset [42]. Designed specifically to address the limitations of conventional benchmarks in crowded scenarios, this dataset serves as a gold-standard platform for evaluating pose estimation under severe mutual occlusion and dense joint clutter. The CrowdPose dataset contains approximately 20,000 images containing roughly 80,000 annotated human instances. Unlike the COCO standard, each human instance in this repository is characterized by a 14-keypoint skeletal

topology tailored for tracking overlapping torsos and limbs. For empirical continuity and strict reproducibility, we followed the standard protocol split, which allocates 10,000 images for training, 2000 images for validation, and 8000 images for testing.

4.2. Experimental environments and hyperparameters

All experiments were conducted on a workstation equipped with a single NVIDIA RTX 4090 GPU. The software environment was configured with Ubuntu 20.04.6, Python 3.10.18, PyTorch 2.8.0, and CUDA 12.8.

The FDC-pose model was trained for 400 epochs with a batch size of 32 and a uniform input resolution of 640×640. We optimized the network using stochastic gradient descent (SGD) with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005. The learning rate was decayed following a cosine annealing schedule to stabilize convergence. Table 1 summarizes the core training hyperparameters.

Table 1. Hyperparameter settings used during training.

Parameter	**Value**
Optimizer	SGD
Momentum	0.937
Weight Decay	0.0005
Initial Learning Rate	0.01
Scheduler	Cosine Annealing
Total Epochs	400
Batch Size	32
Image Size	640×640
Hardware	NVIDIA RTX 4090

4.3. Evaluation metrics

This section introduces the metrics used to evaluate the performance of the FDC-pose model in terms of object detection and HPE. These metrics provide a comprehensive understanding of the model's accuracy and architectural efficiency.

4.3.1. Object keypoint similarity

For the HPE task, we primarily employ the object keypoint similarity metric. OKS quantifies the similarity between predicted keypoints and corresponding ground truth keypoints by calculating the normalized distance between them. The OKS formula is expressed as:

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{4.1}$$

where d_i represents the Euclidean distance between the predicted and ground truth keypoint, s denotes the object scale, k_i is a specific keypoint constant adjusted for object size, and v_i is the visibility flag for keypoint i.

4.3.2. Average precision

The area under the precision-recall curve serves as the central metric for object detection performance, assessing detection quality across different recall levels. The mean average precision (mAP) is calculated as follows:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i, \tag{4.2}$$

where n is the total number of object classes, and AP_i represents the AP for the i-th class. The following AP metrics are reported:

1) **AP:** The primary metric representing the average precision over IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05.

2) **AP_{50}:** Average precision at an IoU threshold of 0.50.

3) **AP_{75}:** Average precision at a stricter IoU threshold of 0.75, assessing high-precision localization.

4.3.3. Model complexity and efficiency

To objectively assess the deployment feasibility of the model on edge-side hardware, this study adopts parameters and giga floating-point operations (GFLOPs) as core evaluation metrics. Parameters (measured in M) reflect the model's static storage requirements and memory footprint, whereas GFLOPs quantify the computational overhead of a single inference, directly determining the inference latency on devices with limited computing power. For real-time HPE tasks, lightweight design (low parameters) alone is insufficient to guarantee real-time performance; an optimal balance between parameter scale and computational load must be sought. This dual optimization strategy aims to break through the computing power bottlenecks of embedded systems, ensuring the algorithm achieves fluid real-time operation while maintaining high precision.

4.3.4. Inference latency

While static parameter scale and theoretical computational complexities establish the mathematical boundaries of network dimensions, physical execution velocity under real-time constraints requires direct temporal quantification. To address this requirement, inference latency is introduced to measure the end-to-end wall-clock execution time consumed by the framework during forward propagation. Expressed consistently in milliseconds per image, this metric determines the processing throughput and operational efficiency of the execution pipeline. For a rigorous evaluation, all latency benchmarks are verified under a standard batch size of 1 executed on an NVIDIA RTX 4090 GPU.

5. Results and analysis

5.1. Experimental results and analysis

We benchmark FDC-pose against diverse HPE paradigms on the COCO 2017 validation set and the CrowdPose dataset. The baselines encompass heavyweight architectures such as deep layer aggregation regression (DLA-reg) [43], lightweight bottom-up models including lightweight

openPose [15] and the efficient series [44], as well as representative single-stage detectors exemplified by super-resolution pose (SRPose) [45], YOLOv8 [46], and YOLOv11 [39]. All evaluations strictly adhere to a unified protocol to ensure fair comparison. To ensure a rigorous and deployment-oriented evaluation, quantitative results on COCO 2017, including AP, parameter count denoted as Params, computational complexity denoted as GFLOPs, and execution latency, are summarized in Table 2.

As shown in Table 2, FDC-pose achieves a highly competitive accuracy-efficiency trade-off on the COCO benchmark. Existing paradigms frequently struggle to balance these multi-dimensional metrics. For instance, SRPose yields an AP of only 48.4% despite requiring 34.1 M parameters and incurring a significant latency of 5.80 ms. The regression-based DLA-reg baseline achieves 51.7% AP but consumes 20.9 M parameters with an execution time of 4.20 ms.

Table 2. Performance comparison of different models on the COCO dataset.

Method	**Params (M)** ↓	**GFLOPs** ↓	**Latency (ms)** ↓	**AP (%)** ↑	**AP_{50} (%)** ↑	**AP_{75} (%)** ↑
DLA-reg	20.9	–	4.20	51.7	81.4	55.2
Lightweight OpenPose	4.1	–	1.35	41.0	66.0	40.6
EfficientHR-H2	10.3	7.7	1.25	52.9	80.5	–
EfficientHR-H3	6.9	4.2	0.95	44.8	76.7	–
EfficientHR-H4	3.7	**2.1**	**0.75**	35.7	69.6	–
SRPose	34.1	30.86	5.80	48.4	–	–
YOLOv8-pose	3.28	9.2	1.15	50.5	80.1	54.3
YOLOv11-pose (Baseline)	**2.86**	7.4	0.95	50.0	80.5	53.5
Ours (FDC-pose)	3.62	9.8	1.85	**53.4**	**83.5**	**57.7**

In contrast, FDC-pose establishes an exceptionally compact yet high-performing architecture. With only 3.62 M parameters (approximately one-sixth of DLA-reg) the model yields 53.4% AP and 83.5% AP_{50}, outperforming DLA-reg by absolute margins of 1.7% and 2.1%, respectively, while cutting the execution time down to 1.85 ms. Furthermore, FDC-pose significantly surpasses lightweight competitors in both precision and velocity, exceeding lightweight OpenPose by 12.4% AP and EfficientHR-H3 by 8.6% AP.

Compared to the YOLOv11 baseline, FDC-pose delivers an absolute AP gain of 3.4%, rising from 50.0% to 53.4%. Notably, the model achieves a 4.2% increase in the stringent AP_{75} metric, improving from 53.5% to 57.7%. Since AP_{75} strictly penalizes minor localization deviations, this substantial gain directly validates the model's efficacy in precisely localizing small-scale targets and heavily occluded instances. By maintaining high consistency in keypoint prediction under partial occlusion, FDC-pose effectively mitigates the cumulative spatial errors common in overlapping target detection.

To explicitly verify the claimed environmental robustness of our framework under severe visual disturbances, we systematically extend our evaluation protocol onto the demanding CrowdPose dataset. This benchmark inherently encapsulates complex real-world scenarios populated by heavily occluded, crowded, and small-scale human targets, serving as an ideal platform to evaluate edge-case resilience. The quantitative cross-domain evaluations are summarized in Table 3.

As demonstrated in Table 3, the AP scores for both DLA-reg and the YOLOv11 baseline on the CrowdPose dataset are 43.9%, while the lightweight configurations experience severe accuracy degradation under dense multi-person clutter.

Table 3. Performance comparison of different models on the CrowdPose dataset.

Method	Params (M) ↓	GFLOPs ↓	Latency (ms) ↓	AP (%) ↑	AP_{50} (%) ↑	AP_{75} (%) ↑
DLA-reg	20.9	–	4.52	43.9	65.6	48.2
Lightweight OpenPose	4.1	–	1.47	32.0	53.6	35.2
EfficientHR-H2	10.3	7.7	1.40	46.5	68.4	–
EfficientHR-H3	6.9	4.2	1.09	36.5	57.4	–
EfficientHR-H4	3.7	**2.1**	**0.86**	26.6	47.1	–
SRPose	34.1	30.86	6.22	41.2	–	–
YOLOv8-pose	3.28	9.2	1.32	43.8	65.8	48.5
YOLOv11-pose (Baseline)	**2.86**	7.4	1.10	43.9	65.7	48.3
Ours (FDC-pose)	3.62	9.8	2.04	**48.4**	**71.6**	**53.8**

In sharp contrast, FDC-pose successfully maintains prominent localization resilience across all precision thresholds, yielding a dominant 48.4% AP, 71.6% AP_{50}, and 53.8% AP_{75}, outperforming the native YOLOv11 baseline by a substantial absolute margin of 4.5% AP. This prominent gain empirically confirms that our contextual dual-branch mechanism and dynamic upsampling successfully decouple tiny, small-scale distal structures from heavy background noise and mutual occlusion boundaries. Consequently, FDC-pose preserves strict topological consistency and mitigates identity-switching defects, thoroughly justifying its claimed robustness under extremely crowded regimes.

Crucially, a joint assessment across both benchmarks underscores the empirical edge and localized stability of our modified network architecture. While FDC-pose outpaces the native YOLOv11-pose baseline by an absolute margin of 3.4% mAP under the standard distributions of the COCO repository, this performance gap scales up to 4.5% mAP upon cross-domain migration into the CrowdPose dataset. This expanding accuracy gap clearly demonstrates our model's superiority in crowded environments. While standard networks often fail when people overlap and limbs are blocked from view, our network successfully identifies individual body structures. Even under severe occlusion, the proposed method accurately aligns the human skeletal joints, proving its strong generalization capability and practical robustness.

In terms of execution efficiency, while the computational overhead scales from 7.4 to 9.8 GFLOPs, FDC-pose registers a pure model inference latency of only 1.85 ms, which corresponds to a processing throughput exceeding 540 frames per second. The nominal latency increment compared to the baseline from 0.95 ms to 1.85 ms is theoretically justified by the operator fragmentation inside the multi-head attention blocks of ODConv and the non-contiguous memory access pattern within the DySample coordinate-mapping interpolation kernel. Crucially, this subtle efficiency investment pays off by successfully rescuing fine-grained distal joints under severe visual disturbances, comprehensively justifying the latency-accuracy trade-off in complex spatial contexts.

To complement the quantitative benchmarks, we visually evaluate FDC-pose under unconstrained scenarios characterized by scale variations, multi-person interactions, and severe occlusion. Figure 8 illustrates comparisons between the YOLOv11 baseline and FDC-pose across varying levels of environmental complexity. Under complex backgrounds and limb overlaps, the baseline suffers from noticeable localization drift at distal joints, alongside missed or false detections. FDC-pose, conversely, effectively infers and reconstructs complete skeletal topologies even when key anatomical parts are partially occluded. By maintaining tight bounding box alignment and precise keypoint

regression amidst severe visual disturbances, FDC-pose effectively preserves topological and structural consistency when handling rare poses and complex occlusion patterns.

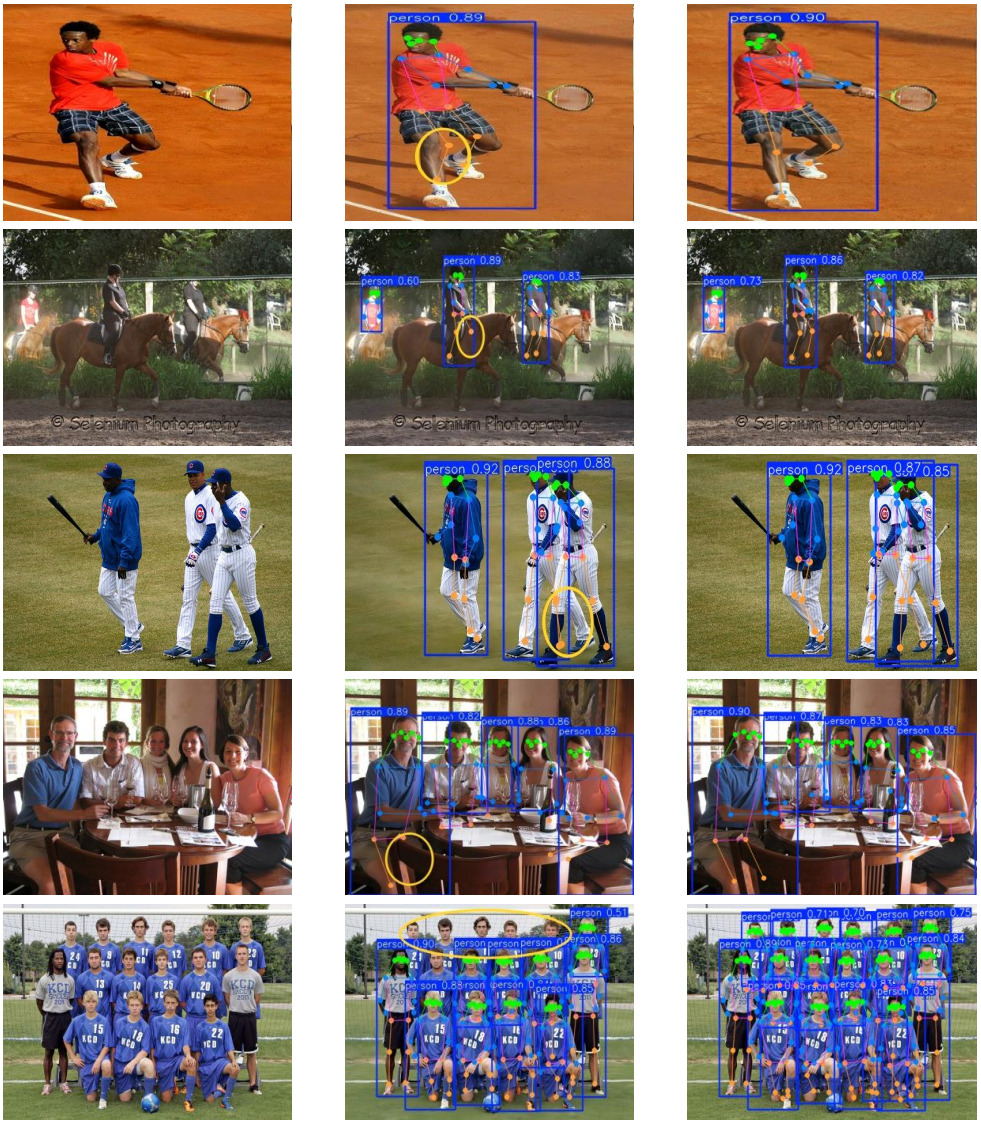


Figure 8. Qualitative comparison of pose estimation results on the validation set under scenarios of increasing complexity. Yellow circles indicate regions with significant estimation errors.

5.2. Ablation study of FDC-pose

We conduct a comprehensive ablation study on the COCO 2017 validation set to systematically evaluate both the individual contributions and the interactive mechanics of our proposed components:

the CPN-FocalOD backbone, the CCFM, the DySample operator, and the MPDIoU loss. Maintained under strictly identical training and evaluation configurations, the quantitative outcomes are categorized into isolated single-component baselines and progressive combinations, as detailed in Table 4.

Table 4. Ablation study of key components in FDC-pose.

Model	CPN-FocalOD	CCFM	DySample	MPDIoU	Params (M) ↓	GFLOPs ↓	AP (%) ↑	AP_{50} (%) ↑	AP_{75} (%) ↑
Baseline					**2.86**	**7.4**	50.0	80.5	53.5
Model A	✓				3.33	8.9	51.1	81.8	54.4
Model B		✓			3.14	8.3	50.6	81.0	54.1
Model C			✓		2.87	**7.4**	50.2	80.7	53.6
Model D				✓	**2.86**	**7.4**	50.3	81.5	53.7
Model E	✓	✓			3.61	9.8	52.7	82.7	57.0
Model F	✓	✓	✓		3.62	9.8	53.2	83.1	57.2
Ours	✓	✓	✓	✓	3.62	9.8	**53.4**	**83.5**	**57.7**

To establish empirical performance baselines for each component, we evaluate the isolated module configurations from Model A through Model D. The baseline YOLOv11-pose model achieves 50.0% AP and 53.5% AP_{75}. Separately replacing the standard backbone with the proposed CPN-FocalOD module, designated as Model A, yields an absolute increase of 1.1% AP and 0.9% AP_{75}, demonstrating that decoupling semantic contexts from dynamic geometric features effectively isolates joint features from background noise. When evaluated in isolation directly on the baseline, introducing the CCFM as Model B or the DySample operator as Model C provides restricted increments of 0.6% and 0.2% AP, respectively. Similarly, applying the MPDIoU loss to the standard baseline without any structural modifications, represented by Model D, results in a minor improvement of 0.3% AP and 0.2% AP_{75}. These single-variable evaluations indicate that while each standalone component positively impacts spatial localization or feature aggregation, their separate contributions remain bounded when facing non-rigid body deformations and severe visual occlusions.

The progressive integration configurations corresponding to Model E, Model F, and ours provide empirical validation of the non-linear synergy mechanism across the multi-stage network pipeline, contesting the assumption of a simple linear plug-in accumulation. If the performance gains were purely additive, the theoretical total precision increment would be bounded by the summation of individual standalone enhancements. Specifically, aggregating the independent improvements of 1.1% from the backbone, 0.6% from the CCFM, 0.2% from the DySample operator, and 0.3% from the loss function yields an expected linear cumulative gain of 2.2% AP, corresponding to a theoretical peak performance of 52.2% AP. However, structural coupling of these components yields a significant non-linear performance increase. Integrating the CPN-FocalOD backbone with the CCFM to form Model E raises the AP to 52.7% and promotes a substantial 3.5% elevation in the strict AP_{75} metric, increasing it from 53.5% to 57.0%. Further incorporation of the content-aware DySample operator as Model F increases the precision to 53.2% AP. Ultimately, the joint orchestration of all modules achieved by our proposed FDC-pose framework secures a peak accuracy of 53.4% AP and 57.7% AP_{75}.

This clear margin between the actual joint improvement of 3.4% AP and the linear additive expectation of 2.2% AP confirms that the multi-stage architecture operates via mutual reinforcement: the backbone filters overlapping limb noise to supply a highly discriminative semantic representation for the neck stage to execute precise point-wise coordinate mapping, while the point-level geometric

boundaries enforced by the MPDIoU loss stabilize bounding box regression and prevent the cascading propagation of spatial localization errors to the keypoint head. This multi-component alignment allows the model to preserve precise topological and structural consistency across rare poses and complex occlusion patterns without requiring supplementary datasets.

Figure 9 illustrates the AP and AP_{50} learning curves throughout the training process, validating the convergence stability of the proposed architecture. The progressive integration of each module induces a consistent upward shift in the performance trajectories. Compared to the baseline, the fully equipped FDC-pose achieves superior peak accuracy and exhibits more stable convergence dynamics in the later training stages, confirming the synergistic robustness of the architectural enhancements.

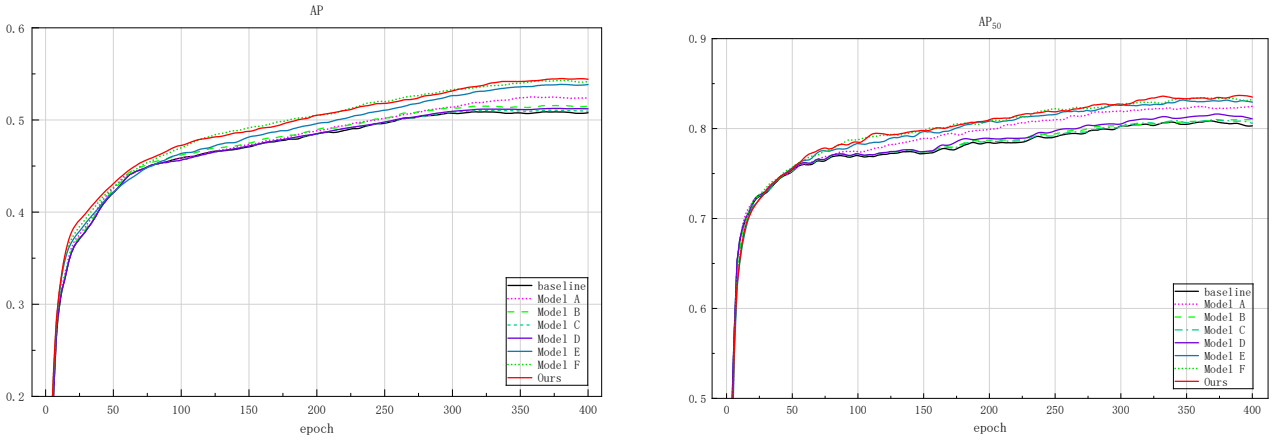


Figure 9. AP and AP_{50} curves throughout the entire iteration process for the model.

Cumulatively, these modules deliver a 3.4% absolute improvement in AP over the baseline. The 4.2% total gain in AP_{75} particularly highlights the model's proficiency in high-fidelity pose estimation. The nominal increase in computational overhead, reaching 3.62 M parameters and 9.8 GFLOPs, is well-justified by the critical localization precision required for rigorous downstream applications.

5.3. Ablation study on backbone architectures

To evaluate the efficacy of the proposed CPN-FocalOD backbone, we benchmark it against six representative architectures. Specifically, the selected baselines encompass the extreme lightweight architecture GhostNetV2 [47], the modern static convolutional network ConvNeXtV2 [48], and the lightweight vision transformer MobileViT [49]. We further expand our evaluation to include the hybrid CNN-transformer EdgeNeXt [50], the highly optimized industrial CNN hierarchical group network V2 (HGNetV2) [38], and the re-parameterization large kernel network (RepLKNet) [51]. The quantitative comparison results are detailed in Table 5.

Substituting the baseline backbone with either GhostNetV2 or ConvNeXtV2 degrades the AP to 48.9% and 49.7%, respectively. This performance decline suggests that the information loss inherent in extreme lightweight linear operations, alongside the purely static spatial nature of ConvNeXt, compromises the fine-grained feature extraction essential for precise keypoint localization.

Table 5. Comparison of different backbone networks on YOLOv11-pose.

Method	Params (M) ↓	GFLOPs ↓	AP (%) ↑	AP_{50} (%) ↑	AP_{75} (%) ↑
YOLOv11-pose (Baseline)	2.86	**7.4**	50.0	80.5	53.5
YOLOv11-pose & GhostNetV2	**2.66**	**7.4**	48.9	80.4	51.8
YOLOv11-pose & ConvNeXtV2	2.82	8.3	49.7	81.1	52.7
YOLOv11-pose & MobileViT-XS	3.45	9.5	50.2	80.8	53.4
YOLOv11-pose & EdgeNeXt-XS	3.48	9.2	50.4	81.0	53.6
YOLOv11-pose & HGNetV2-Tiny	3.52	9.8	50.7	81.3	53.9
YOLOv11-pose & RepLKNet	3.57	10.4	50.6	81.1	53.8
YOLOv11-pose & CPN-FocalOD (Ours)	3.33	9.8	**51.1**	**81.8**	**54.4**

When evaluating models within a comparable parameter scale of 3.4 to 3.6 M, MobileViT-XS and EdgeNeXt-XS yield marginal AP improvements, reaching 50.2% and 50.4%, respectively. Although self-attention mechanisms enhance global context modeling, they inherently lack the fine-grained spatial precision required to handle severe non-rigid pose deformations. Conversely, HGNetV2-Tiny and RepLKNet demonstrate stronger competitiveness by achieving AP scores of 50.7% and 50.6%, thereby underscoring the substantial benefits of optimized hierarchical features and expanded receptive fields.

The proposed CPN-FocalOD outperforms all evaluated architectures, achieving a peak AP of 51.1% and an AP_{75} of 54.4% with a computational overhead of only 3.33 M parameters and 9.8 GFLOPs. By integrating omni-dimensional dynamic filtering, this module effectively decouples overlapping geometric features and adapts to spatial input variations. These results confirm that the performance gains of FDC-pose stem from superior feature disentanglement rather than a mere increase in model capacity, establishing an optimal balance between parameter efficiency and localization fidelity.

5.4. Ablation study on bounding box regression loss

To comprehensively evaluate the efficacy of the MPDIoU bounding box regression loss, we benchmark it against eight established metrics within the YOLOv11 framework. Specifically, our comparative analysis encompasses the standard CIoU baseline alongside early distance-aware approaches, namely GIoU [30] and DIoU. Furthermore, we extend our evaluation to include more recent structural and geometric advancements: the generalized power variant α-IoU [52], the shape-and-scale-focused Shape-IoU [53], the angle-aware scylla intersection over union (SIoU) [54], the aspect-ratio-driven efficient intersection over union (EIoU) [55], and the dynamic non-monotonic strategy denoted as wise intersection over union (WIoU) [56]. The quantitative outcomes of these comparisons are detailed in Table 6.

As detailed in Table 6, the CIoU baseline establishes an AP of 50.0%. Early metrics, namely GIoU and DIoU, underperform this baseline by yielding AP scores of 49.5% and 49.8%, respectively. This performance gap underscores the necessity for comprehensive geometric constraints. Similarly, both EIoU and WIoU result in performance degradation, failing to surpass the 50.0% AP threshold. Such a decline implies that explicit aspect ratio constraints or complex dynamic weighting strategies might introduce optimization instability when applied to the highly non-rigid topological structures of human keypoints. Conversely, α-IoU, Shape-IoU, and SIoU demonstrate competitive performance with AP scores ranging from 50.1% to 50.2%, thereby validating the utility of adaptive weighting and

directional alignment.

Table 6. Comparison of different bounding box regression loss functions on YOLOv11-pose.

Method	AP (%) ↑	AP_{50} (%) ↑	AP_{75} (%) ↑
YOLOv11-pose & CIoU (Baseline)	50.0	80.5	53.5
YOLOv11-pose & GIoU	49.5	80.1	52.8
YOLOv11-pose & DIoU	49.8	80.3	53.2
YOLOv11-pose & EIoU	49.8	80.4	52.9
YOLOv11-pose & WIoU	49.9	80.9	53.2
YOLOv11-pose & α-IoU	50.1	80.8	53.5
YOLOv11-pose & Shape-IoU	50.1	81.0	53.4
YOLOv11-pose & SIoU	50.2	81.2	53.6
YOLOv11-pose & MPDIoU (Ours)	**50.3**	**81.5**	**53.7**

Ultimately, the adopted MPDIoU achieves the highest accuracy across all metrics, reaching 50.3% AP and 53.7% AP_{75}. By directly minimizing the normalized squared Euclidean distance between corresponding bounding box corners, MPDIoU enforces a stricter geometric alignment than standard global shape attributes. This explicit point-level constraint proves significantly more robust to the intrinsic deformations of human bodies, ensuring tight keypoint-to-anatomy alignment in complex occluded scenes.

6. Discussion

The experimental results validate the efficacy of FDC-pose in handling non-rigid deformations and severe occlusions across structurally distinct environments. Compared to the YOLOv11 baseline and mainstream lightweight models, FDC-pose demonstrates superior disentanglement of complex overlapping limbs. Yielding an AP_{50} of 83.5% and an overall AP of 53.4% on the COCO benchmark, alongside a dominant 48.4% mAP on the CrowdPose dataset, the CPN-FocalOD backbone effectively guides the network to focus on visible joints while inferring occluded topological structures. This multi-stage mechanism significantly reduces false positives and missed detections in crowded environments.

Module-level analysis reveals the critical roles of the CCFM and DySample operators in high-precision localization. The 4.2% improvement in AP_{75} on COCO and 53.8% AP_{75} on CrowdPose confirm that these components successfully suppress the detail degradation common in standard feature pyramids. By replacing static convolutions and nearest-neighbor interpolation with a synergistic content-aware strategy, FDC-pose strictly aligns semantic contexts with geometric details, effectively mitigating localization drift for small distal joints. Additionally, the MPDIoU loss imposes stringent point-level geometric constraints, ensuring regression robustness against aspect ratio distortions. This explicit spatial alignment is highly advantageous for precise skeletal reconstruction in real-world scenarios.

Crucially, comparing the results on both datasets clearly highlights the practical advantage of our model in crowded scenes. On the standard COCO dataset, FDC-pose outperforms the standard YOLOv11 baseline by 3.4% AP. When tested on the much more difficult CrowdPose dataset where

people heavily overlap, our model's advantage increases to 4.5% AP. This growing gap directly proves that as the environment becomes more crowded, our dual-branch design works effectively to protect the body structure from being confused by overlapping individuals, instead of failing under severe occlusions.

Despite these advancements, we openly point out the specific conditions under which our model might fail, in order to guide future updates:

First, for the CPN-FocalOD module, although its attention mechanism handles normal body blocking well, its accuracy drops in extremely packed scenes where people overlap by more than 80%. In such crowded cases, the severe background noise confuses the global average pooling layer. This confusion leads the ODConv branch to calculate incorrect attention weights (a_{si}, a_{wi}), which occasionally causes the joint tracking to drift.

Second, although the MPDIoU loss helps stabilize the bounding box corners, it creates a training problem at the very beginning. When the predicted box and the ground-truth box are too far apart and do not overlap at all, the distance penalty term ($d_1^2 + d_2^2$) can become too large and dominate the total loss. This imbalance forces the model to focus too much on fixing the box position rather than learning the keypoints, causing training instability or sudden gradient spikes in the first few epochs.

Third, for the upsampling part, although DySample helps keep edges sharp, it creates small positioning errors when handling very small joints, such as ankles in low-resolution images that take up less than 4×4 pixels. In these low-resolution cases, the learned sampling offsets (O) change too sharply, causing the pixel mapping to shake or become unstable. This instability creates small visual errors that stop the high-precision AP_{75} score from rising further.

Future work will focus on fixing these issues by introducing gradient clipping to balance the loss function, adding a smoothness rule to limit the DySample offsets, and using cross-frame tracking to keep the movements smooth in videos.

7. Conclusions

This study presents FDC-pose, an efficient single-stage framework explicitly optimized to resolve non-rigid deformations and severe occlusions in HPE. The architecture achieves substantial performance gains through three core innovations: 1) the CPN-FocalOD backbone, which synergizes global context modulation with omni-dimensional dynamic filtering to spatially calibrate kernel weights and mitigate feature ambiguity; 2) the integration of the CCFM and the DySample operator, which employs content-aware upsampling to reconstruct fine-grained structural details and accurately localize minute distal joints; and 3) the MPDIoU loss, which enforces strict point-level geometric constraints to stabilize bounding box regression under scale and aspect ratio variations. Extensive evaluations on the standard COCO dataset demonstrate that FDC-pose achieves highly competitive performance, yielding 53.4% AP, 83.5% AP_{50}, and 57.7% in the stringent AP_{75} metric. Symmetrical evaluations on the demanding CrowdPose dataset further confirm its environmental robustness, yielding a prominent 48.4% AP, 71.6% AP_{50}, and 53.8% AP_{75} along with an absolute gain of 4.5% AP over the baseline. Qualitative assessments further confirm its efficacy in resolving complex overlapping poses compared to the baseline. Ultimately, FDC-pose significantly improves the trade-off between high-fidelity spatial localization and computational efficiency, establishing it as a robust solution for demanding applications such as human action analysis, sports evaluation, and

affective computing.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions, which have significantly improved the quality of this manuscript. This work was supported by the Research Startup Fund of Fujian University of Technology (Grant No. GY-Z21064).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
2. C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, et al., Deep learning-based human pose estimation: A survey, *ACM Comput. Surv.*, **56** (2023), 1–37. https://doi.org/10.1145/3603618
3. M. Ma, X. He, X. Bai, A survey of deep learning-based human pose estimation: Methods, datasets, and evaluation metrics, in *2025 2nd International Conference on Digital Image Processing and Computer Applications (DIPCA)*, (2025), 146–152. https://doi.org/10.1109/DIPCA65051.2025.11042554
4. Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, preprint, arXiv:2107.12213.
5. Y. Wang, P. Liu, H. Kang, D. Wu, D. Miao, ICFNet: Interactive-complementary fusion network for monocular 3D human pose estimation, *Neurocomputing*, **616** (2025), 128947. https://doi.org/10.1016/j.neucom.2024.128947
6. A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in *Computer Vision-ECCV 2016*, **9912** (2016), 483–499. https://doi.org/10.1007/978-3-319-46484-8_29
7. B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in *Computer Vision-ECCV 2018*, **11210** (2018), 472–487. https://doi.org/10.1007/978-3-030-01231-1_29
8. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7103–7112. https://doi.org/10.1109/CVPR.2018.00742

9. K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5686–5696. https://doi.org/10.1109/CVPR.2019.00584
10. Y. Xu, J. Zhang, Q. Zhang, D. Tao, ViTPose: Simple vision transformer baselines for human pose estimation, preprint, arXiv:2204.12484.
11. D. Yang, Y. Ge, N. Xu, R. Shi, DDCEFormer: Dual-domain cross enhanced transformer for 3D human pose estimation, *Neurocomputing*, **681** (2026), 133333. https://doi.org/10.1016/j.neucom.2026.133333
12. C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, et al., Lite-HRNet: A lightweight high-resolution network, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 10435–10445. https://doi.org/10.1109/CVPR46437.2021.01030
13. Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, *Comput. Vision Image Understanding*, **192** (2020), 102897. https://doi.org/10.1016/j.cviu.2019.102897
14. Z. Geng, K. Sun, B. Xiao, Z. Zhang, J. Wang, Bottom-up human pose estimation via disentangled keypoint regression, preprint, arXiv:2104.02300.
15. Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh, OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257
16. C. Wang, A. Bochkovskiy, H. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
17. T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, et al., RTMPose: Real-time multi-person pose estimation based on MMPose, preprint, arXiv:2303.07399.
18. Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic convolution: Attention over convolution kernels, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11027–11036. https://doi.org/10.1109/CVPR42600.2020.01104
19. C. Li, A. Zhou, A. Yao, Omni-dimensional dynamic convolution, preprint, arXiv:2209.07947.
20. J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, D. Lin, CARAFE: Content-aware reassembly of features, preprint, arXiv:1905.02188.
21. S. Liu, D. Huang, Y. Wang, Learning spatial fusion for single-shot object detection, preprint, arXiv:1911.09516.
22. Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, et al., Enhancing geometric factors in model learning and inference for object detection and instance segmentation, *IEEE Trans. Cybern.*, **52** (2022), 8574–8586. https://doi.org/10.1109/TCYB.2021.3095305
23. G. Lan, Y. Wu, F. Hu, Q. Hao, Vision-based human pose estimation via deep learning: A survey, *IEEE Trans. Hum.-Mach. Syst.*, **53** (2023), 253–268. https://doi.org/10.1109/THMS.2022.3219242
24. D. Maji, S. Nagori, M. Mathew, D. Poddar, YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss, preprint, arXiv:2204.06806.

25. J. Chen, X. Wang, Z. Guo, X. Zhang, J. Sun, Dynamic region-aware convolution, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 8060–8069. https://doi.org/10.1109/CVPR46437.2021.00797
26. J. Yang, C. Li, X. Dai, L. Yuan, J. Gao, Focal modulation networks, preprint, arXiv:2203.11926.
27. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 8759–8768. https://doi.org/10.1109/CVPR.2018.00913
28. S. Huang, Z. Lu, R. Cheng, C. He, FaPN: Feature-aligned pyramid network for dense image prediction, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 844–853. https://doi.org/10.1109/ICCV48922.2021.00090
29. W. Liu, H. Lu, H. Fu, Z. Cao, Learning to upsample by learning to sample, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2023), 6004–6014. https://doi.org/10.1109/ICCV51070.2023.00554
30. H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 658–666. https://doi.org/10.1109/CVPR.2019.00075
31. X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, et al., Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, preprint, arXiv:2006.04388.
32. S. Ma, Y. Xu, MPDIoU: A loss for efficient and accurate bounding box regression, preprint, arXiv:2307.07662.
33. J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, et al., Human pose regression with residual log-likelihood estimation, preprint, arXiv:2107.11291.
34. Y. Ma, L. Xu, Y. Zhang, T. Zhang, X. Luo, Steganalysis feature selection with multidimensional evaluation and dynamic threshold allocation, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024), 1954–1969. https://doi.org/10.1109/TCSVT.2023.3295364
35. Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167
36. G. Huang, Y. Sun, Z. Liu, D. Sedra, K. Q. Weinberger, Deep networks with stochastic depth, in *Computer Vision-ECCV 2016*, **9908** (2016), 646–661. https://doi.org/10.1007/978-3-319-46493-0_39
37. Y. Ma, L. Xu, Q. Zhang, Y. Zhang, X. Xin, X. Luo, EIS-OBEA: Enhanced image steganalysis via opposition-based evolutionary algorithm, *IEEE Trans. Inf. Forensics Secur.*, **20** (2025), 3616–3631. https://doi.org/10.1109/TIFS.2025.3549692
38. Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, et al., DETRs beat YOLOs on real-time object detection, preprint, arXiv:2304.08069.
39. R. Khanam, M. Hussain, YOLOv11: An overview of the key architectural enhancements, preprint, arXiv:2410.17725.

40. H. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, et al., AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 7157–7173. https://doi.org/10.1109/TPAMI.2022.3222784

41. T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, et al., Microsoft COCO: Common objects in context, preprint, arXiv:1405.0312.

42. J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, C. Lu, CrowdPose: Efficient crowded scenes pose estimation and a new benchmark, preprint, arXiv:1812.00324.

43. F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 2403–2412. https://doi.org/10.1109/CVPR.2018.00255

44. C. Neff, A. Sheth, S. Furgurson, J. Middleton, H. Tabkhi, EfficientHRNet: Efficient and scalable high-resolution networks for real-time multi-person 2D human pose estimation, *J. Real-Time Image Process.*, **18** (2021), 1037–1049. https://doi.org/10.1007/s11554-021-01132-9

45. H. Wang, J. Liu, J. Tang, G. Wu, Lightweight super-resolution head for human pose estimation, in *Proceedings of the 31st ACM International Conference on Multimedia*, (2023), 2353–2361. https://doi.org/10.1145/3581783.3612236

46. J. Terven, D. Córdova-Esparza, J. Romero-González, A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS, *Mach. Learn. Knowl. Extr.*, **5** (2023), 1680–1716. https://doi.org/10.3390/make5040083

47. Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, Y. Wang, GhostNetV2: Enhance cheap operation with long-range attention, preprint, arXiv:2211.12905.

48. S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, et al., ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 16133–16142. https://doi.org/10.1109/CVPR52729.2023.01548

49. S. Mehta, M. Rastegari, MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer, preprint, arXiv:2110.02178.

50. M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, R. M. Anwer, et al., EdgeNeXt: Efficiently amalgamated CNN-transformer architecture for mobile vision applications, preprint, arXiv:2206.10589.

51. X. Ding, X. Zhang, J. Han, G. Ding, Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 11953–11965. https://doi.org/10.1109/CVPR52688.2022.01166

52. J. He, S. Erfani, X. Ma, J. Bailey, Y. Chi, X. Hua, Alpha-IoU: A family of power intersection over union losses for bounding box regression, preprint, arXiv:2110.13675.

53. H. Zhang, S. Zhang, Shape-IoU: More accurate metric considering bounding box shape and scale, preprint, arXiv:2312.17663.

54. Z. Gevorgyan, SIoU loss: More powerful learning for bounding box regression, preprint, arXiv:2205.12740.

55. Y. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, T. Tan, Focal and efficient IOU loss for accurate bounding box regression, preprint, arXiv:2101.08158.

56. Z. Tong, Y. Chen, Z. Xu, R. Yu, Wise-IoU: Bounding box regression loss with dynamic focusing mechanism, preprint, arXiv:2301.10051.

AIMS Press