Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhanced Probabilistic 2D Human Pose Estimation via Adaptive Probability Map and Multi-scale Context Fusion

Jun Zhu
Jinling Institute of Technology
Zhongyuan Xu
Jinling Institute of Technology
Lei Feng
fenglei4923272780126.com

Jinling Institute of Technology
Hao Li
Jinling Institute of Technology
Jiahao Dai
Jinling Institute of Technology
Jiwei Xu
Jinling Institute of Technology

Research Article

Keywords: 2D Human Pose Estimation, Probabilistic Modeling, Adaptive Probability Map, Multi-scale Context Fusion, Image Keypoint Localization

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9812927/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Neural Processing Letters on October 26th, 2026. See the published version at https://doi.org/10.1007/s11063-026-11890-x.

Enhanced Probabilistic 2D Human Pose Estimation via Adaptive Probability Map and Multi-scale Context Fusion

Jun Zhu[1], Zhongyuan Xu[2], Lei Feng[3*], Hao Li[4], Jiahao Dai[5], Jiwei Xu[6]

[1*]School of Computer Engineering, Jinling Institute of Technology, Hongjing Avenue, Nanjing, 211199, Jiangsu, China.
[2]School of Software Engineering, Jinling Institute of Technology, Hongjing Avenue, Nanjing, 211199, Jiangsu, China.
[3]School of Computer Engineering, Jinling Institute of Technology, Hongjing Avenue, Nanjing, 211199, Jiangsu, China.
[4]School of Software Engineering, Jinling Institute of Technology, Hongjing Avenue, Nanjing, 211199, Jiangsu, China.
[5]School of Software Engineering, Jinling Institute of Technology, Hongjing Avenue, Nanjing, 211199, Jiangsu, China.
[6]School of Software Engineering, Jinling Institute of Technology, Hongjing Avenue, Nanjing, 211199, Jiangsu, China.

*Corresponding author(s). E-mail(s): fenglei492327278@126.com;
Contributing authors: zhujun@jit.edu.cn; 19511326180@163.com;
15706123026@163.com; daijiahaoc@163.com; xjw1776375560@126.com;

Abstract

In recent years, human pose estimation has made significant progress in computer vision. However, existing methods still face two major challenges. First, heatmap-based deterministic representations are insufficient for modeling the uncertainty of keypoint predictions and cannot effectively handle out-of-image keypoints. Second, single-scale feature modeling fails to capture both local details and global structural information, resulting in limited performance under occlusion and truncation scenarios. To address these issues, we propose a probabilistic modeling-based framework for 2D human pose estimation. Specifically, we build upon the ProbPose framework and introduce a Multi-Scale Context Fusion

Vision Transformer (MSCF-ViT) as the backbone to enable effective fusion of multi-scale features through cross-scale interactions. Furthermore, an Adaptive Probability Map (APM) module is designed to dynamically adjust spatial probability distributions based on keypoint types and occlusion conditions, enhancing uncertainty modeling capability. To further improve performance, we propose an Adaptive OKSLoss (A-OKSLoss) that aligns the training objective with the evaluation metric while strengthening the learning of hard samples. Extensive experiments on COCO2017, CropCOCO, and OCHuman datasets demonstrate that the proposed method outperforms existing approaches, especially in challenging scenarios involving occlusion and out-of-bound keypoints.

Keywords: 2D Human Pose Estimation, Probabilistic Modeling, Adaptive Probability Map, Multi-scale Context Fusion,Image Keypoint Localization

1 Introduction

2D human pose estimation is a fundamental task in computer vision, aiming to localize the coordinates of 17 human body keypoints from a single image. It serves as a core technology for various high-level applications[1, 2], including action recognition, intelligent surveillance, human–computer interaction, and virtual avatars in the metaverse. Current mainstream top-down[3–7] pose estimation methods are primarily based on heatmap regression, which has become the dominant paradigm due to its stability and localization accuracy. However, these methods still suffer from two inherent limitations in real-world scenarios. First, both training and evaluation focus only on visible keypoints within the image, completely ignoring those located outside the image boundary, leading to frequent mismatches and missed detections in cropped or close-up images. Second, traditional heatmaps[8] rely on fixed Gaussian distributions and are optimized using mean squared error (MSE), without probabilistic calibration. As a result, they cannot model prediction uncertainty or adapt well to occlusion, complex backgrounds, and truncation, thus limiting generalization performance.

ProbPose[9]introduces a fully probabilistic framework into pose estimation by replacing traditional heatmaps with normalized probability maps and incorporating a keypoint existence prediction branch, enabling explicit modeling of out-of-image keypoints. It also constructs the CropCOCO dataset and proposes the Extended OKS (Ex-OKS) metric to evaluate such scenarios. Despite these advances, ProbPose still faces limitations in complex environments. Its probability maps lack adaptability, as they are constrained only by pixel-wise OKSLoss[10] without incorporating priors such as keypoint type or occlusion status, making it difficult to handle diverse keypoint characteristics. Moreover, its single-scale Vision Transformer[11–14] backbone cannot effectively capture both fine-grained local details and global structural context, limiting performance under occlusion and truncation. Additionally, the dual-heatmap strategy increases receptive field but fails to enable effective feature interaction between branches, leading to a trade-off between large receptive fields and localization precision.

To address these issues, this paper proposes an enhanced probabilistic pose estimation framework that is lightweight and easy to implement. While preserving the core design of normalized probability maps and existence prediction, improvements are made in feature modeling, probability map generation, and loss optimization. Specifically, we propose a Multi-Scale Context Fusion Vision Transformer (MSCF-ViT) backbone to integrate global structure and local details through multi-scale embeddings and cross-scale interactions. We further design an Adaptive Probability Map (APM) module that dynamically adjusts spatial distributions based on keypoint types and occlusion conditions, enabling more flexible modeling. In addition, we introduce an Adaptive OKSLoss (A-OKSLoss) and refine the keypoint existence branch to better emphasize hard samples and align training with evaluation.

Extensive experiments on COCO2017[15], CropCOCO[9], and OCHuman[16] datasets demonstrate that the proposed method significantly outperforms state-of-the-art approaches such as HRFormer-s, ViTPose-s, and ProbPose-s in terms of mAP and Ex-mAP. Notably, it achieves substantial improvements in scenarios involving out-of-image keypoints and heavy occlusion. All proposed enhancements are lightweight, easy to reproduce, and do not rely on complex multi-branch refinement structures, maintaining the efficiency of ProbPose while substantially improving robustness and generalization in real-world scenarios.

2 Related Work

In this section, we briefly review the research closely related to our proposed method, which can be categorized into three aspects: mainstream 2D human pose estimation methods, probabilistic pose estimation, and multi-scale contextual feature modeling.

2.1 Mainstream Methods for 2D Human Pose Estimation

Current 2D human pose estimation methods can be broadly categorized into top-down[3–7], bottom-up[17, 18] and single-stage[19–22] approaches. Among them, top-down methods have become the dominant paradigm in the research community due to their superior accuracy. These methods first employ a human detector to localize individual person bounding boxes and then perform keypoint localization within each bounding box. The core approaches can be further divided into two paradigms: heatmap regression and coordinate regression.

Heatmap-based methods represent keypoints using Gaussian heatmaps with the same spatial resolution as the input image. Owing to their strong robustness and ease of training, they have become the mainstream solution. For instance, HRNet[4] maintains high-resolution representations through parallel multi-resolution branches, enabling continuous multi-scale feature fusion and serving as a benchmark for convolutional architectures. ViTPose[6] introduces a pure Vision Transformer architecture into pose estimation, leveraging its capability to model long-range dependencies and achieving state-of-the-art performance on the COCO dataset. However, these traditional heatmap-based methods rely on fixed Gaussian label distributions and are optimized using mean squared error (MSE), leading to uncalibrated outputs that fail to model uncertainty. Moreover, they completely ignore out-of-image keypoints.

In contrast, coordinate regression methods directly predict the 2D coordinates of keypoints, eliminating the need for heatmap decoding and thus providing faster inference. DeepPose[23] first introduced end-to-end regression into pose estimation, while RLE[24] improves regression modeling through residual log-likelihood estimation. Nevertheless, these methods generally underperform heatmap-based approaches in terms of localization accuracy and robustness, especially in scenarios involving occlusion and truncation.

2.2 Multi-scale Context Feature Modeling Based on Vision Transformer

The Vision Transformer[25–30]is a key technique for improving the robustness of pose estimation models. Its core idea is to capture both global human body topology and fine-grained local joint details through features with different receptive fields, thereby overcoming the limitations of single-scale representations.

HRFormer[7] combines high-resolution convolutional representations with Transformer architectures to enable parallel multi-scale feature modeling. Swin Transformer[3] introduces a hierarchical design with shifted window attention, facilitating multi-scale feature fusion. However, these methods are primarily designed for traditional heatmap-based paradigms and are not optimized for probabilistic pose estimation, thus failing to provide sufficient global contextual support for reasoning about out-of-image keypoints.

2.3 Probabilistic Human Pose Estimation

Probabilistic modeling has recently emerged as a promising direction in pose estimation, aiming to transform deterministic keypoint localization into interpretable probabilistic inference. This paradigm addresses the limitations of traditional heatmap methods, particularly the lack of calibration and the inability to represent uncertainty.

ProbPose represents a milestone in this direction by introducing a comprehensive probabilistic pose estimation framework. Specifically, it replaces conventional Gaussian heatmaps with normalized probability maps generated via Sparsemax[31] activation, ensuring valid probability distributions. In addition, it incorporates a keypoint existence prediction branch to explicitly model out-of-image keypoints. In addition, it further optimizes the pixel-wise OKSLoss to better align the training objective with evaluation metrics. Furthermore, the CropCOCO dataset and the Extended OKS (Ex-OKS) metric are introduced, filling the gap in evaluating out-of-image keypoints.

3 Model Design

3.1 Overall Framework

The proposed method adopts an end-to-end top-down architecture, following a hierarchical pipeline consisting of input preprocessing, multi-scale feature extraction, adaptive keypoint probability prediction, output decoding, and supervised learning, as illustrated in Figure 1.

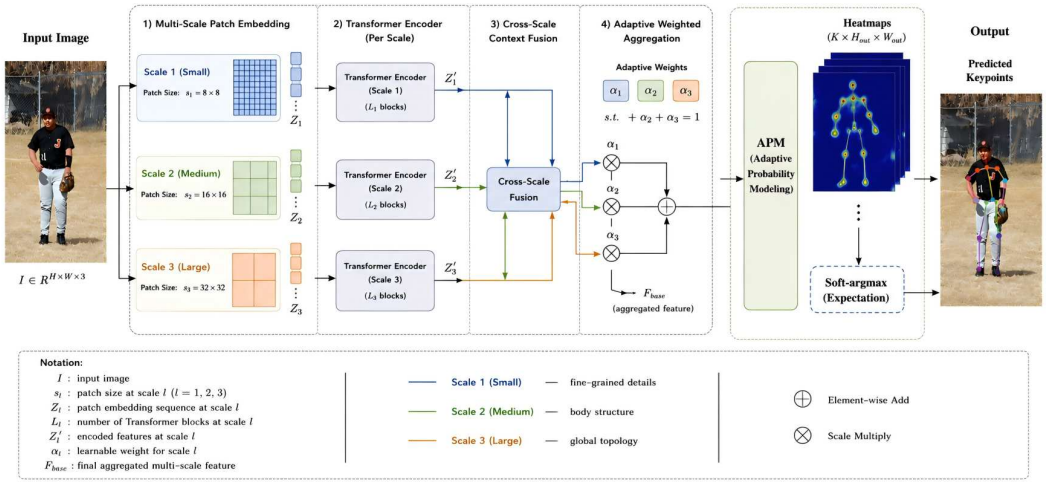


Fig. 1 Overall framework

Given a normalized cropped human image as input, the model first employs the Multi-Scale Context Fusion Vision Transformer (MSCF-ViT) backbone to extract both fine-grained local details and global human topology, producing aggregated multi-scale deep features. These features are then fed into two branches: the Adaptive Probability Map (APM) module and an enhanced keypoint existence prediction branch. The former generates normalized adaptive probability maps for each keypoint, while the latter predicts whether a keypoint lies within the activation window. Finally, keypoint coordinates are decoded via an expected OKS maximization strategy, and the model is optimized end-to-end using ground-truth annotations and the proposed adaptive OKSLoss. By preserving the probabilistic modeling paradigm of ProbPose, the proposed framework enhances both feature representation and probability distribution modeling through MSCF-ViT and APM, achieving improved accuracy in standard scenarios and stronger robustness in challenging conditions.

3.2 Multi-Scale Context Fusion ViT Backbone (MSCF-ViT)

The proposed MSCF-ViT serves as the fundamental feature extractor, designed to overcome the limitations of the single-scale Vision Transformer used in ProbPose, including insufficient multi-scale perception, weak local detail modeling, and limited global reasoning capability.

Given an input image I, we first divide it into three groups of patches with different sizes, corresponding to small, medium, and large receptive fields. The multi-scale patch embeddings are formulated as:

$$Z_l = \text{PatchEmbed}_{s_l}(I), \quad l = 1, 2, 3 \tag{1}$$

where s_l denotes the patch size at scale l, and Z_l represents the corresponding feature sequence.

Each scale is processed independently by a Transformer encoder to model spatial dependencies:

$$\mathrm{Attn}(Z_l) = \mathrm{Softmax}\left(\frac{Q_l K_l^\top}{\sqrt{d}}\right) V_l \tag{2}$$

To enable cross-scale interaction, features from all scales are concatenated and fused:

$$F_{\text{fusion}} = \mathrm{Conv}_{1\times1}\big(\mathrm{Concat}(Z_1, Z_2, Z_3)\big) \tag{3}$$

The fused features are then added back to each scale via residual connections:

$$Z_l' = \mathrm{Attn}(Z_l) + F_{\text{fusion}} \tag{4}$$

Finally, the multi-scale features are aggregated using adaptive weighting:

$$\mathcal{F}_{\text{base}} = \sum_{l=1}^{3} \alpha_l Z_l', \quad \text{s.t.} \sum_{l=1}^{3} \alpha_l = 1 \tag{5}$$

where α_l are learnable weights. This design enables effective integration of global structure and local details, providing strong contextual features for subsequent probability map generation.

3.3 Adaptive Probability Map (APM) Module

In ProbPose, probability maps are generated using a fixed formulation, which, although effective for basic keypoint localization, lacks adaptability to different keypoint types, occlusion levels, and scene complexities. Such fixed distributions fail to capture structural differences among facial, torso, and limb keypoints, and cannot provide differentiated responses for visible, occluded, and out-of-image keypoints.

To address these limitations, we propose an Adaptive Probability Map (APM) module as illustrated in Figure 2., which replaces fixed distributions with adaptively calibrated probability maps. The APM leverages multi-scale contextual features extracted by MSCF-ViT to perform spatial calibration while preserving the normalization constraint via Sparsemax.

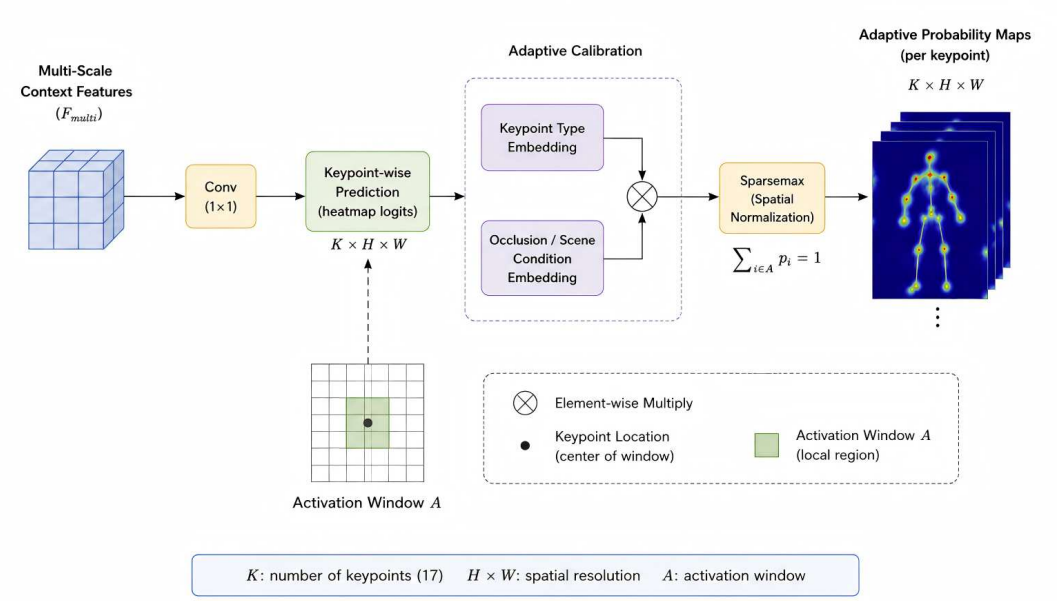


Fig. 2 Adaptive Probability Map

Formally, for the j-th keypoint, the adaptive localization probability at pixel x_i is defined as:

$$p_L^A(x_i) = p\big(x_i \mid k_j \in \mathcal{A}, \mathbf{I}, \mathcal{F}_{\text{multi}}\big) \tag{6}$$

where x_i denotes the pixel coordinate, k_j represents the j-th keypoint, $\mathcal{A}$ denotes the activation window, $\mathbf{I}$ is the input image, and $\mathcal{F}_{\text{multi}}$ corresponds to the multi-scale contextual features extracted by MSCF-ViT.

The probability map is normalized within the activation window using Sparsemax:

$$\sum_{x_i \in \mathcal{A}} p\big(x_i \mid k_j \in \mathcal{A}, \mathbf{I}, \mathcal{F}_{\text{multi}}\big) = 1 \tag{7}$$

To further enhance adaptability, the APM dynamically adjusts spatial responses according to keypoint type and occlusion status. Specifically, a context-aware modulation function is introduced:

$$\tilde{p}(x_i) = \phi(k_j, s) \cdot p_L^A(x_i) \tag{8}$$

where $\phi(k_j, s)$ denotes an adaptive modulation factor conditioned on the keypoint type k_j (e.g., face, torso, limb) and scene complexity s (e.g., occlusion level or out-of-image status).

The resulting adaptive probability map provides a more flexible and expressive representation of keypoint localization, enabling the model to assign more reasonable probability distributions under diverse conditions. This significantly improves robustness in challenging scenarios such as occlusion, truncation, and out-of-image keypoints.

3.4 Loss Function

Based on the adaptive localization probability $p_L^A(x_i)$, this paper proposes an Adaptive OKSLoss (A-OKSLoss). Built upon the pixel-level OKSLoss proposed by ProbPose, this loss introduces an adaptive weighting mechanism[3] to further align the training objective with the evaluation metric, and strengthens the supervision constraint on hard sample keypoints such as occluded and out-of-image keypoints.

The A-OKSLoss takes minimizing the expected risk as the optimization objective. First, we define the adaptive expected risk $R_{\text{exp}}^A(x_i)$:

$$R_{\text{exp}}^A(x_i) = w(k_j; s) \cdot \big(1 - \text{OKS}(x_i)\big) \cdot p_L^A(x_i) \tag{9}$$

where $w(k_j; s)$ is the adaptive weight, dynamically generated by the APM module according to the keypoint type k_j (face/trunk/limb) and scene complexity s (occlusion degree, whether the keypoint is outside the activation window), which is used to increase the loss weight of hard-to-localize keypoints; $\text{OKS}(x_i)$ is the Object Keypoint Similarity value of the target keypoint corresponding to pixel x_i; $p_L^A(x_i)$ is the adaptive localization probability .

On this basis, the pixel-level adaptive OKSLoss is defined as:

$$\mathcal{L}_{\text{A-OKS}}(x_i) = (1 - \alpha)\, R_{\text{exp}}^A(x_i) + \alpha\, g(x_i) \tag{10}$$

where $R_{\text{exp}}^A(x_i)$ is the aforementioned adaptive expected risk; $g(x_i)$ is the heatmap gradient term, calculated by the Sobel operator, which acts as a regularization term to constrain the smoothness of the probability map; the hyperparameter α controls the regularization strength, and its value is consistent with that of ProbPose.

Compared with the original OKSLoss, A-OKSLoss assigns higher optimization weights to hard samples such as occluded and out-of-image keypoints through the adaptive weight $w(k_j; s)$, guiding the model to focus on the keypoint localization accuracy in complex scenes. Meanwhile, it retains the Sparsemax activation and gradient regularization constraints, which not only ensures that the probability distribution of each keypoint satisfies the normalization requirement of summing to 1 within the activation window, but also minimizes the probability difference between adjacent pixels through smooth regularization, avoiding sharp peaks and overfitting in the probability map, thus obtaining a more robust probability distribution.

In the inference phase, this paper still adopts the decoding strategy based on expected OKS maximization: calculate the expected OKS value for each pixel in the activation window and take the argmax, then optimize the sub-pixel localization accuracy through quadratic interpolation. Relying on the adaptive localization probability, this strategy further improves the localization performance in complex scenes such

as bimodal heatmaps and occluded keypoints, outperforming the traditional direct argmax decoding method.

4 Experiments

4.1 Experimental Setup

4.1.1 Experimental Environment

All experiments are conducted under a unified hardware and software environment. The hardware platform adopts NVIDIA RTX 4090 GPU with 24GB video memory and NVIDIA RTX 3090 Ti GPU with the same video memory specification. The CPU is Intel Xeon E5-2650 V4 (12 cores, 24 threads, base frequency 3.0GHz, turbo frequency 4.1GHz), which can quickly respond to GPU computing requirements and collaboratively process complex computing tasks. The software framework is built based on the PyTorch deep learning framework and MMPose pose estimation toolbox, combined with CUDA 12.1 and cuDNN 8.9 acceleration operators to achieve efficient forward inference and backpropagation, ensuring the stability and efficiency of the model training and testing process. The experimental datasets cover standard pose datasets and special datasets for complex scenes, including the COCO2017 dataset for conventional pose evaluation, the CropCOCO dataset for out-of-image keypoint verification, and the OCHuman dataset for severe occlusion scene evaluation, which can comprehensively verify the effectiveness and robustness of the model under different conditions.

4.1.2 Experimental Datasets

This paper conducts experiments on the following three public datasets:

- **COCO 2017**[15]: The standard benchmark dataset in the field of 2D human pose estimation, containing 118,287 training images and 5,000 validation images, annotated with 17 human joint keypoints.
- **CropCOCO**[9]: A cropped dataset constructed based on the COCO2017 validation set, containing 4,114 images where some human keypoints are located outside the activation window, specifically designed to evaluate the model's ability to localize out-of-image keypoints.
- **OCHuman**[16]: A special pose estimation dataset for severe occlusion scenes, containing 5,081 highly occluded human images and 13,360 human annotations, compatible with the COCO 17-keypoint annotation specification, used to verify the model's anti-interference ability and localization robustness under limb occlusion and human overlap scenes.

4.1.3 Implementation Details

The proposed method is implemented based on the MMPose framework and PyTorch. To ensure fair comparison, the training hyperparameters are basically consistent with ProbPose: the training batch size is 64, the optimizer is AdamW (weight decay coefficient 0.1), the initial learning rate is 5×10^{-4} (linearly scaled according to the batch

size), and the total number of training epochs is 210. Learning rate warm-up is performed for the first 500 steps, and the learning rate is decayed to 0.1 times the original at the 170th and 200th epochs. The input image size is 256×192, the heatmap output size is 64×48, and the input image padding coefficient is 1.25. Horizontal flip data augmentation is adopted in the testing phase to improve localization accuracy.

4.1.4 Comparison Methods

This paper compares the proposed method with current mainstream 2D human pose estimation methods, including:

- Traditional heatmap-based methods:HRNet-W32[4], HRFormer-s[7],PVTv2[32], Swin-T[3], ViTPose-s[6], TokenPose-s[33]
- Regression-based methods: RLE[24]
- Probabilistic methods: ProbPose-s[9], ProbPose-s-DH[9] (the dual heatmap version of ProbPose)

To ensure fair comparison, all comparison methods adopt the same backbone network size and training hyperparameters, and uniformly use manually annotated human bounding boxes during testing.

4.1.5 Evaluation Metrics

The core evaluation metrics include: mean Average Precision (mAP), precision under different IoU thresholds (AP.5, AP.75), precision for different scale targets (AP(M), AP(L)), Average Recall (AR), and Extended mean Average Precision (Ex-mAP) specifically designed to evaluate the localization performance of out-of-image keypoints.

4.2 Main Experimental Results

4.2.1 Experimental Results on COCO2017 Dataset

Table 1 shows the performance comparison of different methods on the COCO2017 validation set. The proposed method achieves the highest mAP and Ex-mAP, with mAP reaching 77.4%, an improvement of 0.8 percentage points over ProbPose-s, and Ex-mAP reaching 76.9%, an improvement of 0.5 percentage points. The gain is mainly reflected in the overall AP and the presence-aware Ex-mAP, while AP(L) and AR remain slightly lower than those of ProbPose-s. This indicates that the proposed modules improve the overall localization distribution without uniformly dominating every sub-metric. The performance improvement mainly comes from three aspects: the MSCF-ViT backbone integrates multi-scale context information and strengthens the representation of global human structure and local joints; the APM module dynamically optimizes the probability map distribution according to keypoint types and scene complexity; and the optimized loss function further aligns the training objective with the evaluation metric.

Table 1 Comparison of keypoint detection accuracy of different pose estimation algorithms on COCO2017 dataset

Method	mAP(%)	AP.5(%)	AP.75(%)	AP(M)(%)	AP(L)(%)	AR(%)	Ex-mAP(%)
HRNet-W32[4]	73.4	89.5	80.7	70.2	80.1	78.9	74.1
HRFormer-s[7]	75.2	90.1	82.3	72.1	81.4	79.6	74.6
PVTv2[32]	72.0	88.1	79.1	68.8	78.6	77.0	71.5
TokenPose-s[33]	75.6	90.3	82.8	72.5	81.8	80.0	75.0
Swin-T[3]	73.5	88.7	80.2	70.3	79.8	78.1	72.9
ViTPose-s[6]	75.9	90.5	83.1	72.8	82.0	80.2	75.3
RLE[24]	74.2	89.5	81.5	71.2	80.7	79.1	73.8
ProbPose-s[9]	76.6	91.0	83.8	73.5	82.6	80.8	76.4
ProbPose-s-DH[9]	76.2	90.8	83.4	73.1	82.2	80.5	75.4
Ours	77.4	93.6	83.8	74.0	81.1	80.7	76.9

4.2.2 Experimental Results on CropCOCO Dataset

Table 2 shows the experimental results of different methods on the CropCOCO dataset, which is specifically designed to evaluate out-of-activation-window keypoint localization. The proposed method achieves the best mAP and Ex-mAP, with mAP reaching 82.1%, an improvement of 0.4 percentage points over ProbPose-s, and Ex-mAP reaching 74.2%, an improvement of 0.3 percentage points. Although AP(M) and AP(L) are marginally lower than those of ProbPose-s, the improvement in mAP, AP.75, AR, and Ex-mAP indicates better overall localization and stronger out-of-window reasoning. This result verifies the effectiveness of the core modules for truncated keypoint localization: the APM module adaptively adjusts the probability map response distribution, and the MSCF-ViT backbone provides context support for structural reasoning of keypoints outside the activation window.

Table 2 Comparison of keypoint detection accuracy of different pose estimation algorithms on CropCOCO dataset

Method	mAP(%)	AP.5(%)	AP.75(%)	AP(M)(%)	AP(L)(%)	AR(%)	Ex-mAP(%)
HRNet-W32[4]	70.5	86.9	78.2	67.0	76.4	75.8	63.9
HRFormer-s[7]	70.9	87.2	78.5	67.3	76.8	76.1	64.3
PVTv2[32]	70.7	86.8	78.0	66.9	76.2	75.6	63.1
TokenPose-s[33]	71.9	88.1	79.6	68.2	77.6	77.0	65.8
Swin-T[3]	71.3	87.5	78.9	67.6	77.1	76.4	65.0
ViTPose-s[6]	72.7	88.6	80.1	68.9	78.3	77.5	66.5
RLE[24]	70.8	87.1	78.6	67.4	76.9	76.2	64.7
ProbPose-s[9]	81.7	93.8	86.2	78.4	84.6	85.2	73.9
ProbPose-s-DH[9]	80.9	93.3	85.7	77.8	84.1	84.7	71.4
Ours	82.1	93.9	86.6	78.2	84.5	85.8	74.2

4.2.3 Experimental Results on OCHuman Dataset

Table 3 shows the performance comparison of different methods on the OCHuman dataset. OCHuman is a challenging pose estimation dataset for severe occlusion, human crowding, limb truncation, and out-of-frame cases, focusing on keypoint perception and structural reasoning under complex interference. The proposed method achieves the best mAP, AP.75, AP(L), AR, and Ex-mAP among the compared methods, demonstrating strong adaptability to complex scenes.

Traditional heatmap-based methods (such as HRFormer-s, Swin-T, ViTPose-s) rely on fixed Gaussian distribution labels and single-scale feature modeling, and their features are prone to failure under conditions of severe occlusion and incomplete human structure, with mAP all below 60.5%. Probabilistic methods ProbPose-s and ProbPose-s-DH improve the fitting ability of occluded keypoints through probability map modeling and uncertainty learning, with mAP reaching 60.4% and 61.4% respectively. However, they are still limited by single-scale features and fixed probability distributions, and have limited reasoning ability for extreme occlusion and human topological structure.

With the synergistic effect of the MSCF-ViT multi-scale context fusion backbone and the APM adaptive probability map module, the proposed method achieves an mAP of 64.0%, improving over ProbPose-s by 3.6 percentage points and over ProbPose-s-DH by 2.6 percentage points. This indicates that multi-scale context fusion can capture global human structure and provide context support for occluded limbs, while the adaptive probability map dynamically optimizes the distribution according to keypoint types and occlusion states. Together, the two modules improve localization robustness in scenes with severe occlusion, human crowding, and limb truncation.

Table 3 Comparison of keypoint detection accuracy of different pose estimation algorithms on OCHuman Dataset

Method	mAP(%)	AP.5(%)	AP.75(%)	AP(M)(%)	AP(L)(%)	AR(%)	Ex-mAP(%)
HRNet-W32[4]	59.8	85.2	68.7	52.2	63.4	66.7	59.3
HRFormer-s[7]	60.3	85.8	69.2	52.7	63.9	67.1	60.0
PVTv2[32]	58.5	84.6	67.3	51.2	62.1	65.8	58.1
TokenPose-s[33]	60.8	86.1	69.6	53.0	64.2	67.4	60.5
Swin-T[3]	58.1	84.3	66.8	50.9	61.7	65.3	57.9
ViTPose-s[6]	60.3	85.9	69.3	52.8	63.9	67.2	60.1
RLE[24]	59.5	85.0	68.5	51.9	63.0	66.5	58.9
ProbPose-s[9]	60.4	86.7	70.2	53.1	64.6	67.3	60.2
ProbPose-s-DH[9]	61.4	85.8	70.1	53.0	65.1	66.9	61.2
Ours	64.0	86.2	70.5	53.1	65.8	67.7	62.3

4.3 Ablation Study

To verify the contribution of the MSCF multi-scale context fusion ViT backbone and the APM adaptive probability map module to the model performance, this paper takes

ProbPose-s as the baseline model and conducts module ablation experiments on the OCHuman dataset. The core modules are added step by step and the metric changes are observed. The experimental results are shown in Table 4. After adding only the MSCF backbone to the baseline model, the model mAP is improved by 1.3 percentage points, proving that multi-scale context fusion can effectively enhance global structure perception and provide more sufficient context reasoning support for occluded and truncated limbs. After adding only the APM module, the model mAP is improved by 1.8 percentage points, indicating that the adaptive probability map can dynamically optimize the distribution according to keypoint occlusion states and types, significantly improving the localization accuracy in complex scenes. When both the MSCF backbone and the APM module are used, the model performance is greatly improved, with mAP reaching 64.0%, an improvement of 3.6 percentage points over the baseline model. This fully confirms that the two core modules are highly complementary and jointly promote the significant improvement of the model's comprehensive performance in difficult scenes such as severe occlusion, human crowding and limb truncation.

Table 4 Module ablation study results

Model Configuration	mAP(%)	AP.5(%)	AP.75(%)	Ex-mAP(%)
ProbPose-s	60.4	86.7	70.2	60.2
ProbPose-s + MSCF	61.7	86.5	70.1	60.8
ProbPose-s + APM	62.2	86.4	70.4	61.6
ProbPose-s + MSCF + APM	64.0	86.2	70.5	62.3

4.4 Fine-grained Analysis and Visualization

4.4.1 Performance Analysis by Keypoint Type

To verify the adaptive optimization effect of the APM module, this paper divides the 17 keypoints of the COCO dataset into three categories: face, trunk and limbs, and counts the AP values of the baseline model ProbPose-s and the proposed method on the CropCOCO dataset respectively. The results are shown in Table 5.

It can be seen from the results that the proposed method improves the AP values on all three types of keypoints, with the AP of face keypoints improved by 0.8 percentage points, the AP of limb keypoints improved by 0.7 percentage points, and the AP of trunk keypoints improved by 0.3 percentage points. This result fully verifies the effectiveness of the APM module: for fine face keypoints, the model adaptively generates a more concentrated probability distribution to improve localization accuracy; for limb keypoints that are prone to occlusion and out-of-frame, the model adaptively optimizes the distribution and strengthens structural reasoning ability, achieving comprehensive optimization of different types of keypoints.

Table 5 AP value comparison of different types of keypoints

Method	Face Keypoints(%)	Trunk Keypoints(%)	Limb Keypoints(%)
ProbPose-s	85.2	89.4	78.6
Ours	86.0	89.7	79.3

4.4.2 Visualization Results

To intuitively verify the robustness of the proposed method, we present qualitative results on COCO, CropCOCO and OCHuman datasets, as shown in Figures 3, 4 and 5. On the COCO dataset (Figure 3), our method achieves stable and accurate keypoint localization in dynamic sports scenes such as tennis and baseball swing, as well as tool interaction scenarios. Benefiting from the strong feature representation ability of the MSCF-ViT backbone and the adaptive probability modeling of the APM module, it effectively handles partial occlusion and large pose deformations, generating complete and physiologically reasonable human skeletons. On the CropCOCO dataset (Figure 4), which focuses on limb truncation and out-of-image keypoints, our method shows obvious advantages over traditional methods that often lead to incomplete or broken skeletons. With the enhanced keypoint existence prediction branch, our model can accurately infer the position distribution of truncated and out-of-frame keypoints, thus maintaining a complete human topological structure. On the OCHuman dataset (Figure 5), our method exhibits strong robustness in challenging scenes with severe occlusion and multi-person interaction. It can effectively distinguish overlapping human bodies, accurately locate heavily occluded keypoints, and avoid skeleton dislocation or disorder that frequently occurs in existing methods. These visualization results are consistent with our quantitative experiments, confirming the effectiveness and generalization ability of the proposed method.



Fig. 3 Visualization results on the COCO dataset



(a) CropCOCO (b) CropCOCO (c) CropCOCO

Fig. 4 Visualization results on the CropCOCO dataset



(a) OCHuman (b) OCHuman (c) OCHuman

Fig. 5 Visualization results on the OCHuman dataset

5 Conclusion

Focusing on the core demand of human keypoint localization in complex scenes, this paper addresses the problems of insufficient utilization of multi-scale context information, fixed probability map generation rules, and limited localization accuracy of out-of-image keypoints in traditional probabilistic pose estimation methods, and proposes a lightweight pose estimation model based on Multi-scale Context Fusion ViT backbone (MSCF-ViT) and Adaptive Probability Map module (APM).

The core contributions and innovations of this paper can be summarized as follows: First, the MSCF-ViT backbone network is designed, which effectively fuses human global topological structure and local joint detail features through multi-scale patch embedding and cross-layer context interaction mechanism, providing high-quality feature support for keypoint localization. Second, the APM adaptive probability map module is proposed, which can dynamically optimize the probability distribution according to keypoint types, occlusion degrees and scene complexity, realize differentiated modeling of in-image and out-of-image keypoints, and significantly improve

the localization robustness in complex scenes. Third, the keypoint existence probability branch and the adaptive OKS loss function are optimized, which assign higher optimization weights to hard samples such as occluded and out-of-image keypoints, achieve end-to-end accurate training, and fully align the training objective with the evaluation metric.

To verify the effectiveness of the model, this paper conducts sufficient comparative experiments and ablation experiments on three datasets: COCO2017, CropCOCO and OCHuman. The experimental results show that the proposed method significantly outperforms existing mainstream methods such as HRFormer-s, ViTPose-s and ProbPose-s on core metrics such as mAP and Ex-mAP, and shows better localization accuracy and generalization ability in both conventional scenes and out-of-image keypoint scenes. The ablation study further confirms that the MSCF-ViT backbone and the APM module have good synergistic effect, and they jointly promote the improvement of model performance, verifying the effectiveness and necessity of each core module.

Although the proposed model achieves good results in the keypoint localization task, it still has certain limitations: First, the localization accuracy of the model in scenes of extreme severe occlusion and dense multi-human overlap still has room for improvement. Second, although the inference speed of the model meets the basic real-time requirements, the lightweight degree for edge device deployment can be further optimized.

Future research directions can be carried out around the following points: First, further optimize the multi-scale feature fusion mechanism and combine attention mechanisms to strengthen feature mining in occluded regions. Second, explore more efficient lightweight network structures to improve the model inference speed while ensuring accuracy, adapting to the real-time deployment requirements of edge devices. Third, extend the proposed method to the field of 3D human pose estimation, and combine multi-view information to achieve more accurate 3D keypoint localization, providing more comprehensive technical support for upper-layer applications such as human action recognition and human-computer interaction.

Declarations

Funding

This work was supported by the Jiangsu University Blue Project.

Conflict of interest

The authors declare no competing interests.

Data availability

The datasets used in this work are publicly available from the COCO, CropCOCO, and OCHuman datasets.

Code availability

Not applicable.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

[1] Zhang, J., Tao, D.: Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things. IEEE Internet of Things Journal **8**(10), 7789–7817 (2020)

[2] Lin, W., Liu, H., Liu, S., Li, Y., Qian, R., Wang, T., Xu, N., Xiong, H., Qi, G.-J., Sebe, N.: Human in Events: A Large-Scale Benchmark for Human-Centric Video Analysis in Complex Events (2020)

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)

[4] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 5693–5703 (2019)

[5] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: PVT v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**, 415–424 (2022)

[6] Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[7] Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: HRFormer: High-resolution transformer for dense prediction. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)

[8] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–499 (2016)

[9] Purkrabek, M., Matas, J.: ProbPose: A probabilistic approach to 2d human pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 27124–27133 (2025)

[10] Maji, D., Nagori, S., Mathew, M., Poddar, D.: YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 2637–2646 (2022)

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2020)

[12] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: DETR: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)

[13] Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7262–7272 (2021)

[14] Pan, Z., Zhuang, B., Liu, J., He, H., Cai, J.: Scalable vision transformers with hierarchical pooling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 377–386 (2021)

[15] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 740–755 (2014)

[16] Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.-M.: Pose2seg: Detection free human instance segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 889–898 (2019)

[17] Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 7291–7299 (2017)

[18] Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 14676–14686 (2021)

[19] Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 11069–11078 (2022)

[20] Stoffl, L., Vidal, M., Mathis, A.: End-to-End Trainable Multi-Instance Pose Estimation with Transformers (2021)

[21] Xiao, Y., Wang, X., Yu, D., Su, K., Jin, L., Song, M., Yan, S., Zhao, J.: AdaptivePose++: A Powerful Single-Stage Network for Multi-Person Pose Regression (2022)

[22] Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 11060–11068 (2022)

[23] Toshev, A., Szegedy, C.: DeepPose: Human pose estimation via deep neural networks. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 1653–1660 (2014)

[24] 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), pp. 11025–11034 (2021)

[25] Xu, Y., Zhang, Q., Zhang, J., Tao, D.: Vitae: Vision transformer advanced by exploring intrinsic inductive bias. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 34 (2021)

[26] Wang, W., Yao, L., Chen, L., Lin, B., Cai, D., He, X., Liu, W.: Crossformer: A versatile vision transformer hinging on cross-scale attention. In: International Conference on Learning Representations (ICLR) (2022)

[27] Zhou, J., Wang, P., Wang, F., Liu, Q., Li, H., Jin, R.: Elsa: Enhanced Local Self-Attention for Vision Transformer (2021)

[28] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 568–578 (2021)

[29] Wang, P., Wang, X., Luo, H., Zhou, J., Zhou, Z., Wang, F., Li, H., Jin, R.: Scaled relu matters for training vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2495–2503 (2022)

[30] Zhang, Q., Xu, Y., Zhang, J., Tao, D.: Vitaev2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond (2022)

[31] Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 1614–1623 (2016)

[32] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)

[33] Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.-T., Zhou, E.: Tokenpose: Learning keypoint tokens for human pose estimation. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)